e-book

Microsoft

Cloud Scale Analytics 101

Unlock the full potential of your data

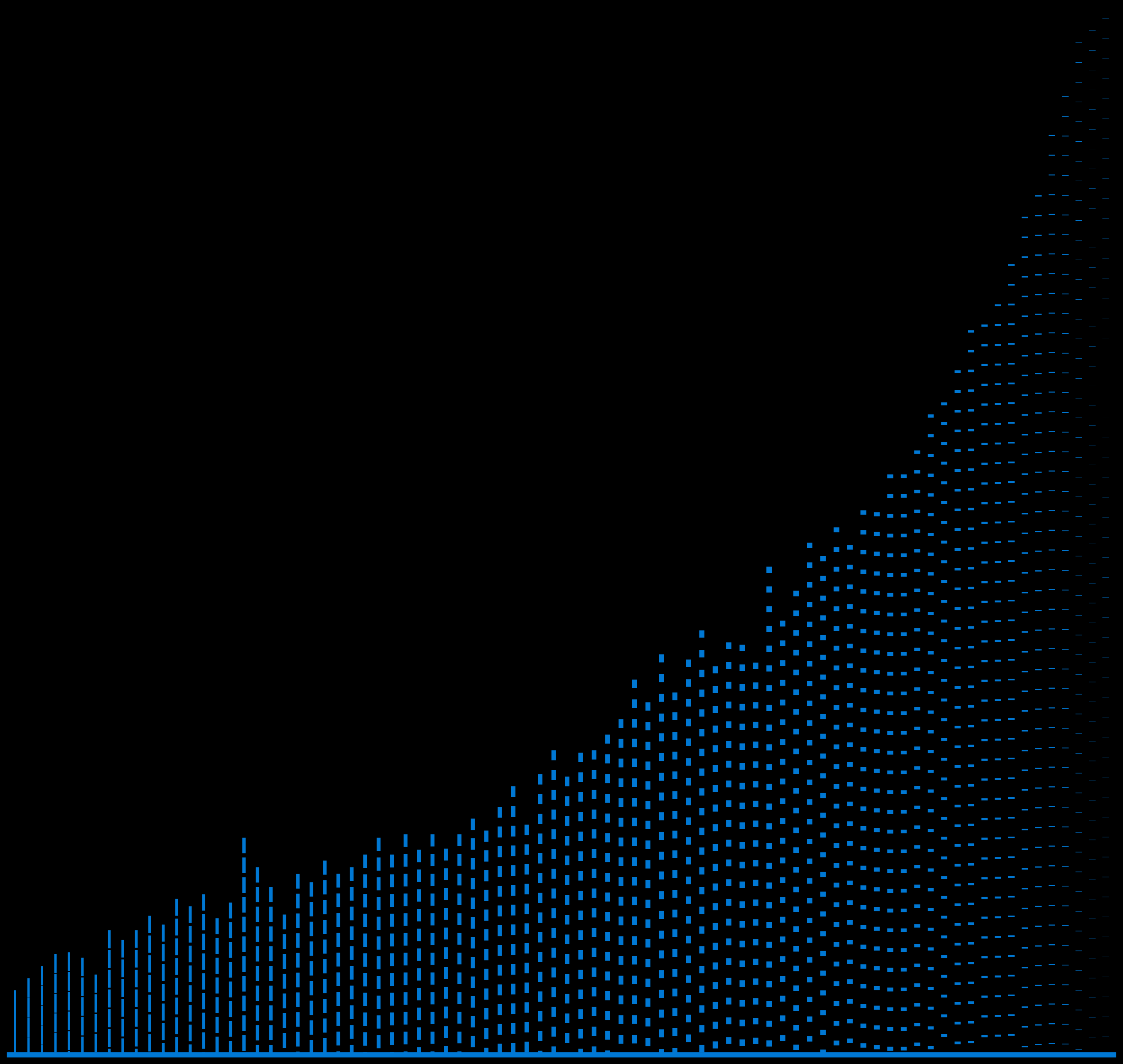


Table of Contents

Introduction

"Analytics" is a word we hear used repeatedly, both in technology and even mainstream business circles. The term is mentioned frequently in many business conversation contexts. But what does it mean and how can it help businesses like yours?

Having a common understanding of what analytics is, and what it can do for you, is vital for business success these days.

In this e-book, we'll explore:

The core concepts of analytics

Appropriate business goals

Underlying technologies that empower you to analyze your data

Analytics can help your organization move to a proactive mode of operations in an age where continuous improvement is the expectation.

What you'll learn:

The What

What is cloud scale analytics and how can it help your business.

The Why

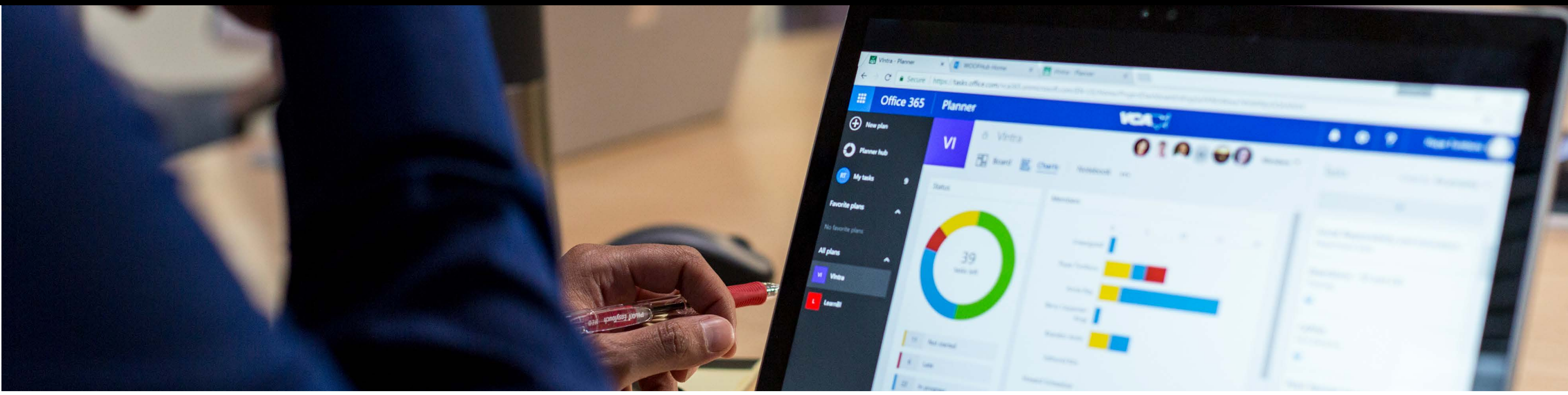
Why a comprehensive cloud scale analytics strategy is critical for your business.

The How

The core concepts and the building blocks that bring cloud scale analytics to life—for you.

Chapter 1
What is cloud scale analytics?

The crux of analytics is based on a distinction between the way software handles data and the way humans do it. Whether it's collecting and storing data or learning from it to improve your strategy, using software to analyze your operations is much more efficient than doing it the old-school way—'by hand'.



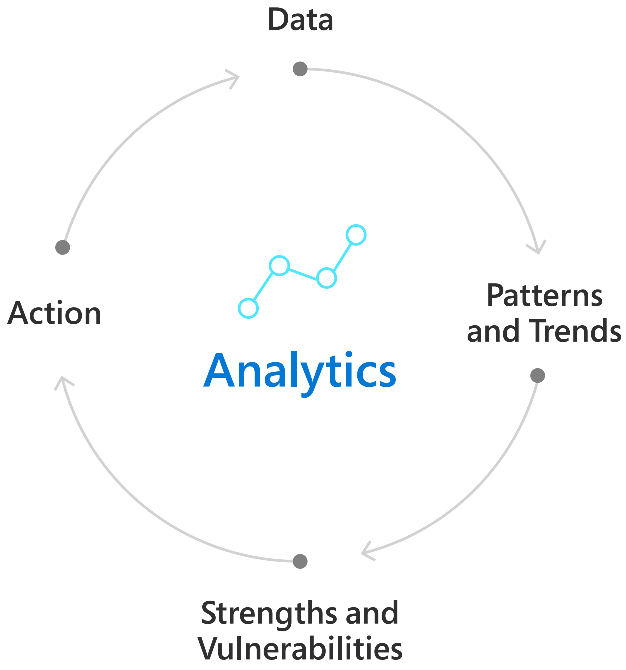
As ethereal as it may sound to get information from data, doing so can be part of your operational norms.

Analytics uses raw data to discern contextual information. Storing data in database tables as a series of neat rows and columns works well to service screens in a line-of-business software application. But handling data in that way is infrastructural work, i.e. the plumbing. And if that's all you do with your data, then arguably most of its value will remain latent, unrealized. Mining that data for information and insight that supports your decision-making is what analytics is all about.

The real hot spot of data is in summarizing it and discerning patterns, trends, and tendencies. This enables your entire organization to understand itself, its strengths, and its vulnerabilities. Armed with this information you can enact initiatives that maximize what you're doing well, and mitigate and improve areas of underperformance. Analytics can help your organization move to a proactive mode of operations in an age where continuous improvement is the expectation.

The secret to outfitting your organization for analytics is coalescing your data from all the many places where it may reside and conforming it structurally so that you can look at it all together, in summarized form.

For example, you can pull records from your invoicing system about individual order line items, and data in your Web analytics system about each click your customers took on your website and when. But you'll realize the power of analytics when you pull all that data (and more) into the same repository and make it compatible to get valuable information such as the total spent per customer and then correlate it with visits to specific areas of your site.



When you get all your data out of its siloed locations, the pieces of the proverbial puzzle start to fit together.

When you get all your data out of its siloed locations, and you organize it so that a given customer's data from one of those silos can be matched up with that same customer's data from another, you begin to get real insight into customer behavior. And when you get all the data in a conformed structure, it becomes easier to look past individual rows of data, towards aggregations of it, and start to observe macro phenomena instead of micro events.

Analytics isn't magic

That may sound rather lofty though…and speaking of analytics in those terms risks abstracting it too much. So, let's be clear here: analytics isn't magic. It's not even really that complex. As ethereal as it may sound to get information from data, doing so can be part of your operational norms.

> When you take analytics to the cloud and do it at scale, the value rises still more.

No more data silos

Now you can discern how sales are trending overall or break that down by groups of customers who visit your website frequently, occasionally, or rarely. If you ingest data from your catalog system, you'll then be able to break that analysis down by product category, or even individual product. By planning carefully, all that trend analysis will help build machine learning models that predict how patterns and trends in your sales will manifest.

That's the value of analytics. But when you take analytics to the cloud and do it at scale, the value rises still more.

In the cloud, as more data is added, your analytics structure will scale, elastically and economically. As the variety of your data increases, the workload flexibility of the cloud will come in especially handy, as each data source may have different levels of structure, and different data volumes that accumulate at different rates of speed. In the cloud, a single service may have the required levels of flexibility, but you'll also have the option of teaming different services together on the same data, because of common platform standards around cloud storage.

You'll have hybrid flexibility too, enabling you to integrate data in your cloud environment with data on-premises, as well as with data from cloud-based Software as a Service (SaaS) applications. And you'll have security and privacy baked in at the platform layer, along with the tools and technologies you'll need to propagate it up the stack into your databases, data lake, and applications.

When you have modern data warehouse, real-time analytics and machine learning technology working together this way, you don't just have the advantages of analytics; you have the efficacy, economies, and the power of cloud scale analytics.

Modern data warehouse + Real-time analytics + Machine learning

Cloud-scale analytics

What it means...

You'll realize the power of analytics when you pull data from different sources and organize them so that customer data from one of the silos can be matched up against that customer's data from another source. Doing so will enable observation of macro trends.

Taking such analytics to the cloud scales the power of data; the possibilities enabled through real time analytics and machine learning create unprecedented scale and insights for action.

Chapter 2
Why cloud scale analytics?

Today, analytics isn't a bonus prize; it's a fundamental requirement. If you don't do the work to coalesce and organize your siloed data, you could be limiting the degree of your success. The roadmap to running your business efficiently lies in your data, and without analytics you're neglecting to use that roadmap and must rely on hunches. Considered in that light, analytics is critical, not optional.

Analytics is critical, not optional.

With analytics, your business is organized, competitive, and optimized.

With analytics, your business is organized, competitive, and optimized and you can make decisions with greater confidence and conviction. That can make your organization more proactive. You'll be more willing to verify your hunches and intuitions and the trepidation you may have around making changes can decrease.

Analytics helps you answer the questions you were afraid to ask. And the fear in asking those questions will dissipate, because you'll have confidence in determining an answer. You won't just find out what things might be going wrong; you'll also figure out how to make them go right.

You've collected and maintained your data. Why not go the last step and analyze it? Insights gained from this analysis can make your organization more efficient, more competitive, more valuable, and more in control. Taken from that perspective, analytics, and especially cloud scale analytics, is a must-do.

Cloud scale analytics is a must-do.

The last step

Again, there's no magic here. The information isn't synthesized from thin air. It already lies dormant in your data. Yes, you need to take steps to emancipate it and, yes, that might seem like extra work. But instead of thinking of such an initiative as an extraordinary effort, think of it for what it is: one of the last steps in making the most of your data efforts, and the phase where most of the value kicks on.

What it means...

The roadmap to running your business efficiently lies in your data and in the insights derived from analytics.

Cloud scale analytics is the phase where most of the value kicks in.

Chapter 3
Sources of data

As we've already discovered, doing analytics the right way requires putting together different overlapping data sets. But doing that correctly requires you to think non-traditionally about the data sources you'll need to mine. Some of the sources are obvious, including all of the databases underlying your operational applications.

Sources of data may include:

- SaaS applications
- Marketing
- Operations
- Finance
- Social Media
- Sales

Whether these are so-called OLTP (online transactional processing) databases running on relational database platforms, or databases running on NoSQL databases, you know that you must get at their data to implement an effective analytics platform.

But there are other important places to go as well. For example, there's lots of important data sitting in files on disk. Many of these may be in a simple format called CSV (comma separated values). The format treats each row in the file as a row of data, with commas acting as the field separator of data within the row.

File-based data is everywhere.

It could be on your own computer. It could be in cloud storage. It could be in "data lake" technologies that we'll discuss in the next chapter. No matter where it is, though, including it in your analytics implementation may be very important.

Along with modern data warehousing and machine learning, real-time insights form a pillar of cloud scale analytics.

Streaming data. Don't forget streaming data, also known as "data in motion." This is data that is constantly being churned out, from websites, on social media, and maybe even from devices, connected to the Internet of Things (IoT). Capturing this data can be a little trickier, but the rewards can be great, as your analytics can be completely up to date, in real-time. Along with modern data warehousing and machine learning, real-time insights form a pillar of cloud scale analytics.

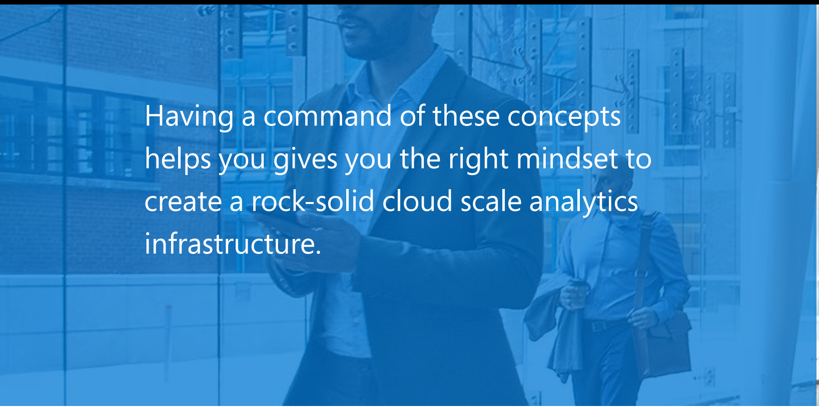
On-premises and cloud. Finally, there are all the applications you're running on-premises and in the cloud. Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), marketing, customer service, web analytics and more. Each of these has an application programming interface (API) that lets you get the data in the application. If the plethora of APIs seems intimidating, consider that connectors exist to help make those APIs look like tables in a database instead.

What it means...

Ensuring that data from all your sources, such as streaming data, on premises and cloud, is available in a cohesive fashion to help provide meaningful cloud-scale analytics implementation.

Chapter 4
Concepts

With a better understanding of where, why, and what (in terms of the types of data sources available) let's look at a few analytics concepts.

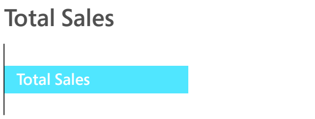


Having a command of these concepts helps you gives you the right mindset to create a rock-solid cloud scale analytics infrastructure.

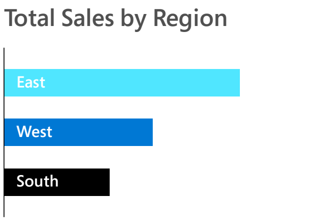
Measures and dimensions

Once you've gathered your data, you'll need to organize it. While leaving the data in its more transactional format is possible, the data you gather will fall into one of two broad categories: the numeric data you want to track, such as sales or page views and the categories you want to break them down by, such as fiscal quarter, product category, or customer.

In the world of analytics, the numeric data points are called measures and the categories are called dimensions. The vocabulary doesn't matter that much though, and these days, the two types of data don't even need to be separated. But to the extent that measures and dimensions can be stored in separate tables, you'll come closer to classic data warehouse design (and, by the way, we'll discuss data warehouses in the next chapter of this e-book).


Total Sales

Example of measure, shown in a chart


Total Sales by Region

Example of measure, broken down by dimensions, shown in a chart

Aggregation approaches

Beyond the structure of your data lies the question of how to aggregate it. In some analytics systems, aggregates (for example, total sales by each combination of customer, product, and calendar date) are precalculated, for optimized performance. But in other analytics systems, they are calculated at the time queries are issued.

Beyond the structure of your data lies the question of how to aggregate it.

While the latter approach can be slower, it also avoids the need to update the pre-calculations, making the analytics system more agile. Beyond that, there are ways to store data which make aggregating it on-demand a faster process. Columnar storage, for example, segregates all the values for each column, making the process of aggregating it much more efficient, since the values for the other columns in each row don't have to skipped over. There are data compression dividends as well, allowing more of the data to fit in memory and be aggregated even more efficiently.

Parallel processing and scale-out architecture

Another technology used to speed up on-demand aggregation is distributed compute architecture, where several computer servers are bundled together in a "cluster." With this approach, when it comes time to query and aggregate the data, each server in the cluster gets a assigned a piece of the work and all servers perform their constituent portions of the query simultaneously.

This divide-and-conquer approach reduces query times. And as data volumes grow, you can add more nodes (servers) to the cluster, to keep the overall query times relatively constant. This parallel processing and the ability to "scale out" (by adding more servers) are hallmarks of both data warehouse and big data systems.

A good analytics platform can handle all data, be it structured, semi-structured, or unstructured.

What about unstructured data...

Revisiting the question of structure, some data, including free text and media (like images and audio) are not very structured at all—in fact, this type of data is referred to as "unstructured." Even CSV files have implied schemas, but data of the types just mentioned do not. Some streaming data, especially data coming from sensors on Internet of Things (IoT) devices, may be semi-structured, where some data is formally organized in columns and other data is unstructured, or varies in structure from row to row.

In addition to IoT streaming data, much of the data from social media may follow this pattern as well. Rigorously structured columns for the date and time of a post may be side-by-side with the raw text of the post or comment itself. A good analytics platform can handle all data, be it structured, semi-structured, or unstructured.

What it means...

An understanding of analytics concepts is essential for making the most of cloud scale analytics.

After getting data, it is important to organize it.

Beyond the structure of the data lies the question of how to aggregate it.

The real-time insights from unstructured and semi-structured data are essential for cloud scale analytics and a good platform accommodates all kinds of data.

Chapter 5
Building blocks

In this section we discuss specific components in the cloud scale analytics arena. We'll describe each of the components broadly as well as identify which products on the Azure cloud platform and in Microsoft's on-premises Enterprise software stack deliver the described functionality.

Your analytics system contains your most valuable business data.

Data warehouse solutions without limits

Data warehouses are analytics-specific databases, built to handle large volumes of data while still employing relational database technology to implement the repository. These days, most data warehouse products use some combination of three technologies:

Columnar storage, described on page 12.

Massively parallel processing (MPP), a clustered, parallel, scale out technology, also described previously.

Vector processing, which enables the central processing unit (CPU) cores on each server to process multiple numerical data values at once, rather than one at a time. This is typically achieved using special "Single Input Multiple Data" (SIMD) CPU instructions.

Azure Synapse Analytics

Adding more nodes to the Azure Synapse Analytics cluster is easy, as is removing nodes. This elasticity allows Azure Synapse Analytics to accommodate customers' immediate needs, by adding nodes when necessary in order to keep up at busier times, and to drop nodes if workload requirements subside. Storage can also be added as needed. Both computing power and storage are independently scalable (Fig 5.2).

The Microsoft Analytics Platform System (APS) belongs to the on-premises side of the product portfolio and was formerly known as SQL Server Parallel Data Warehouse (PDW). It supports many of the same features as Azure Synapse Analytics, but lacks the SSD caching layer and, being based on a physical appliance, is sized more statically than Synapse Analytics.

Fig. 5.1

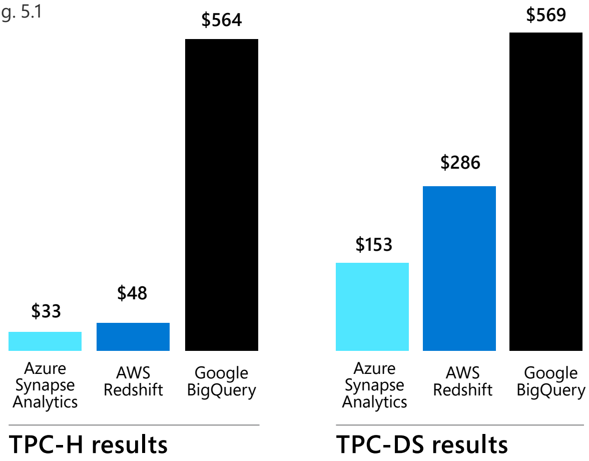
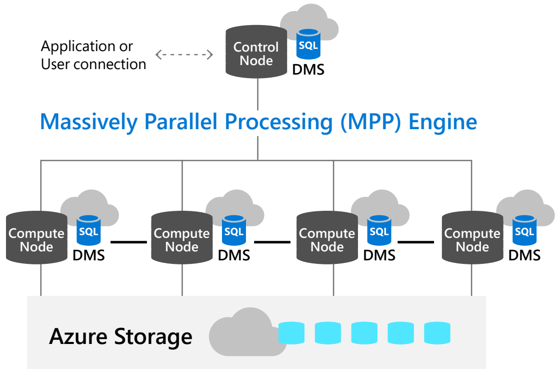


Fig. 5.2



This elasticity allows Azure Synapse Analytics to accommodate customers' immediate needs.

Azure Synapse Analytics (Synapse Analytics) is Microsoft's cloud contender in the data warehouse category, shown in recent independently conducted benchmarks (Fig 5.1) to beat out the competition on price and performance, sometimes dramatically. While based on the same technology that underlies the OLTP-oriented Azure SQL Database (and the on-premises SQL Server product), Azure Synapse Analytics sports an MPP architecture, provides columnar storage through a feature called columnstore indexes, and implements vector processing through its Batch Mode feature. It also accelerates performance by using a solid-state drive (SSD) storage caching layer over the underlying cloud storage persistence layer.

Organizations can use HDInsight to avail themselves of a data lake sweet spot.

Azure Synapse Analytics effectively demonstrates the power of cloud scale analytics.

In addition to its elasticity, built-in security, and data masking privacy features, as well as its membership in the broader SQL Server ecosystem, Synapse Analytics also features the ability to **pause and resume** all compute resources, adding significant economic efficiency. This allows compute-oriented expenses to be mitigated, sometimes significantly, while the affordability of durable cloud storage means no data is lost.

Azure Synapse Analytics excels at analytics workloads on high volumes of data that is well structured, partially summarized, and relevant to the entire enterprise. Synapse Analytics's versatility extends beyond such canonical DW workloads, though. For example, Synapse Analytics can also process semi-structured data formatted as JSON (or XML). Moreover, it can handle the kind of granular data often associated with OLTP databases, with new indexing features that let Synapse Analytics perform "needle in the haystack" searches to find transactional records.

Other new capabilities allow Synapse Analytics to handle multiple workloads, in multiple clusters, with varying SLAs, to be implemented within a single Synapse Analytics instance, making it a platform that serves both data warehouse and data mart requirements.

Data lake technologies

We've already discussed how files on disk are bona fide data sources. If that piqued your curiosity, you'll be interested in data lake technology. Data lakes let you leave data in file form, using formats such as CSV, or a newer yet very popular columnar format called Apache Parquet.

Azure Databricks and Azure HDInsight are scale-out data lake platforms capable of numerous workloads, including streaming big data analytics, data engineering, machine learning, and streaming data processing. Azure Databricks is based on Apache Spark, and though the product was developed by Databricks (the company founded by Spark's creators), it is offered and supported as a first-party service from Microsoft.

Azure HDInsight

Azure HDInsight uses Apache Hadoop and its YARN cluster resource manager to host a variety of open source analytics projects, including Apache Hive for SQL access, Pig for data transformation, HBase for NoSQL workloads, Storm and Kafka for streaming data processing. It offers Microsoft Machine Learning Services for data science and artificial intelligence, and it hosts an open source implementation of Apache Spark.

Business Intelligence (BI) allows for fast analysis of a variety of data.

When it comes to data lakes, there's also the power of the ecosystem, with the HDInsight application platform. Throughout this chapter, we'll mention examples of where this ecosystem partner program can integrate products and services from third parties that make HDInsight-based data lakes even more powerful.

Organizations can use HDInsight to choose one or a combination of technologies to organize and analyze an array of data sets kept in cloud storage. The combination of Hadoop, Spark, Hive, Kafka, HBase and other open source engines makes HDInsight a perfect platform for agile analytics on high volumes of semi-structured data.

Azure Databricks

Azure Databricks is a fantastic platform for customers who want the fastest implementation of Spark available, to perform a combination of data engineering, streaming data processing, data analytics and machine learning workloads, in mix-and-match fashion, on unstructured and semi-structured data. In the on-premises world, the forthcoming SQL Server 2019 will offer its own data lake/Big Data solution based on a combination of Hadoop, Spark and SQL Server itself. By allowing SQL Server to query the same data in the same Hadoop Distributed File System storage layer as Apache Spark, Microsoft offers a full choice of processing engines and paradigms with which to perform data lake analytics.

Business Intelligence

Business Intelligence (BI) allows for fast analysis of a variety of data, most of it structured. BI platforms can build columnar storage repositories around data that has been organized into measures and dimensions and can quickly analyze it. Some BI systems focus on data visualizations, reports, and dashboards while others are focused on curated back-end repositories, optimized for analytical queries. Some platforms focus on both.

Power BI is a versatile platform for data visualization and exploration at the personal, team and Enterprise levels.

Microsoft Power BI

Microsoft Power BI offers industry leading data visualization and a dizzying array of data connectors, allowing for the construction of BI models geared to lightning-fast in-memory query performance. Power BI Desktop is a free end-user application, but the platform works best when combined with the Power BI cloud service. The latter is available in three types of subscriptions: free, Professional, or Premium which provide individual, departmental, or Enterprise capabilities, respectively. With premium subscriptions, subscribers get a dedicated infrastructure, and the ability to scale it out to multiple servers and unlimited consumption users.

The Power BI engine is based on the technology found in the on-premises SQL Server Analysis Services platform or Azure Analysis Services, a standalone cloud service. Both Power BI and Azure Analysis Services focus on Analysis Services' more modern Tabular mode, a columnar BI engine, rather than its older Multidimensional mode which is based on OLAP (online analytical processing) technology. The SQL Server Analysis Services platform supports both modes. And because the engine technologies are common, models built in both of those modes can be queried by Power BI, allowing for a powerful hybrid solution of cloud-based BI reports and dashboards and on-premises BI back-end infrastructure.

Power BI excels at supporting direct connectivity to back-end data sources at analysis time, using a technology called DirectQuery, available in a great many of its data source connectors. And while users have long had the choice of DirectQuery or the standard Import models, Power BI now offers a composite model option, allowing users to mix and match. This can work especially well in concert with platforms we've already discussed, such as Azure Synapse Analytics, HDInsight, and Databricks, where aggregated data can be kept in the import portion of a composite model and the voluminous detail data these systems are great at managing can be queried and aggregated via DirectQuery.

Power BI is a versatile platform for data visualization and exploration at the personal, team, and Enterprise levels. It pairs extremely well with data in Synapse Analytics and HDInsight as well as raw data in Azure Storage (Blob Storage and Azure Data Lake Storage – details below) and transactional data in Azure SQL DB, and SQL Server. In fact, Power BI can connect to data in virtually any Microsoft data platform technology, both in the cloud and on-premises, as well as to a huge array of non-Microsoft data sources.

Only cloud scale analytics, with its workload flexibility, hybrid integration capabilities and leverage of ecosystem power can bring all this bear.

Data Virtualization and Hybridizing Technologies

The question of whether to use Import, DirectQuery or Composite models in Power BI relates to a concept we investigated at the beginning of this e-book: coalescing and conforming siloed data in order to analyze it. But while that process of coalescing may seem to imply physical movement of the data, that is not necessarily the case. While Power BI DirectQuery demonstrates this, there are other technologies that can both leave the data in place and allow you to treat it logically as if it were local.

Such technologies fall under the category of **data virtualization**, which is growing in popularity because it is so useful. As data volumes grow and the number of data sources grows as well, physical movement and transformation of data from every source becomes prohibitive. While physical movement of selected data can improve performance dramatically, starting with a virtualized data baseline saves time and money. It's also a great way to ensure security and privacy, as data virtualization platforms can both manage role-based access and respect the access controls in place on individual data sources. Leaving the data in place leaves data source security intact as well.

PolyBase technology

Microsoft's PolyBase technology provides data virtualization services for Azure Synapse Analytics, as well as on-premises SQL Server, and Analytics Platform System. It supports external tables where the metadata is kept locally and actual data remains in place at the source. Developers working against the database, however, can treat external tables and standard tables equivalently and can even join tables of each type in a single query. PolyBase is a great way to integrate data in Azure Blob Storage and Azure Data Lake Storage logically into Synapse Analytics, SQL Server, and APS, enabling query of that data with the Transact SQL (T-SQL) language.

PolyBase also works on Cloudera and Hortonworks Hadoop clusters and will soon work against Oracle, Teradata, MongoDB, and even other SQL Server instances, as well as any ODBC-compatible data source. In addition to PolyBase, Azure Data Lake Analytics offers a SQL query interface directly over data in Azure Blob Storage and ADLS. Further, the HDInsight Application

In combination with cloud object storage and on-premises distributed storage, such files, or groups of them, can form the foundation for a basic data lake.

Platform makes available third-party solutions like Starburst Presto, which provides its own data virtualization platform with an MPP query engine as the interface.

Use PolyBase when you want to analyze data in the lake, while leveraging your team's T-SQL skill set on SQL Server, Azure SQL DB and/or Synapse Analytics, combining external data with that stored in the database or warehouse.

Only cloud scale analytics, with its workload flexibility, hybrid integration capabilities, and leverage of ecosystem power can bring all this bear.

Storage is a cloud scale technology too

Microsoft's Azure Data Lake Storage (ADLS which is based on Azure Blob Storage) is a great medium for data lakes. It's built to handle files of arbitrarily large size and supports a true hierarchical (folder-based) file system. This is important in the data lake world, since very often a large group of files, stored in the same folder or folder-subfolder hierarchy, needs to be treated as a single data set. The supported folder-level operations in ADLS facilitate this.

All of us are used to conventional storage on our computers (i.e. hard disks, or solid-state drives) and many of us are familiar with shared Enterprise storage. But now the data, storage, and cloud worlds have aligned to drive the twin trends of object storage and distributed storage.

Object storage

Object storage is the name of the game in the cloud; its container paradigm has acquainted users with the concept of economical storage that is fully elastic, according to point-in-time needs. Azure Blob Storage is Microsoft's cloud object storage offering. As we've seen, ADLS builds on top of it, to transcend size limits in individual files, containers, and accounts, as well as to provide hierarchical file system services that support folder-wise operations.

Distributed storage

The Hadoop Distributed File System (HDFS) provides something similar in the on-premises world, by aggregating the conventional disks on each server in a cluster to form a logically unified, but physically distributed, file system. Like cloud object storage, HDFS supports elastic expandability. It also supports fault tolerance and resiliency through maintaining multiple replicas of the files it manages, such that failure of any one node in the cluster will not result in data loss. That is why SQL Server 2019 Big Data Clusters, which can operate on-premises, utilizes HDFS in its storage pools.

ADLS is the common thread between most components of the Microsoft advanced analytics stack.

We already mentioned that the simple file formats like CSV can be used in data lake scenarios. In combination with cloud object storage and on premises distributed storage, such files, or groups of them, can form the foundation for a basic data lake.

Parquet

But what really makes the data lake paradigm compelling is to go beyond such simple "flat" file storage and work with more sophisticated file formats, that are optimized for analytical querying scenarios. We briefly mentioned the Parquet file format previously. Like the slats of wood in its namesake floor tiles, Parquet files store data in columnar fashion. Normally we'd think of columnar storage as being in the realm of data warehouse and BI technology, but Parquet brings this chunk of database engine technology into the storage and data lake worlds. In so doing, it slightly blurs the distinction between data lake and data warehouse platforms (technologically speaking) and brings greater efficacy to the former.

Not only is Parquet columnar, but it supports quite granular partitioning, at the file and folder level, which can make certain analytical queries (for example, those based on a specific time period) even more efficient. Spark, Hive and Databricks Delta can read Parquet files natively, making the format especially well-suited to Azure Databricks and HDInsight. The SQL Server 2019 database engine will also have platform-level compatibility with the Parquet file format, bringing it into relevancy for relational database-skilled professionals as well.

The combination of ADLS and Parquet is the perfect storage solution when your data volumes are large or are expected to grow, and you want to access all data from a variety of services. ADLS is the common thread between most components of the Microsoft advanced analytics stack, including Azure Databricks, HDInsight, Azure Data Factory, and Power BI. With your data in ADLS, you can process it with any one or combination of those and other services.

Manage and protect your data to ensure sensitive data is only visible to the people with appropriate permissions.

Data integration

Despite the convenience and elegance offered by data virtualization technology, at least some data will need to be physically transformed, in order to be conformed to, and properly integrated with, data from other sources. Getting the transformation work done will require the use of data engineering, data preparation, and data pipeline technology.

PowerBI. The Microsoft stack offers a rich array of such technologies. Power BI Desktop's built-in Power Query tool offers sophisticated data preparation and profiling capabilities. Power BI dataflows, a new feature of Power BI Premium, ports many of PowerQuery's capabilities (including its underlying programming language, called "M") to the cloud.

Azure Data Factory and Databricks. On the Azure platform, Azure Data Factory offers sophisticated data pipelining capabilities. And because Spark is extremely well-suited to very sophisticated data engineering jobs, Azure Databricks is a great platform for data integration too.

HDInsight. HDInsight works well for this purpose as well, for a couple of reasons. HDInsight supports Spark, giving you access to the data engineering sophistication of that platform. But HDInsight supports other platforms including Apache Pig, which, along with its "Pig Latin" language, is geared specifically to data transformation. In addition, the HDInsight Application Platform supports third-party data preparation and data engineering products like Trifacta and Datameer, putting ecosystem power to work once more.

Azure Data Factory. Azure Data Factory (ADF) works well for all types of data processing functions, be it extract transform and load (ETL), extract load and transform (ELT), or simple data ingest and movement. The addition of Mapping Data Flows (now in public preview) and Wrangling Data Flows (now in private preview) enables ADF to be a no-code/low-code platform for running data engineering workloads on Azure Databricks.

HDInsight is a great solution in data lake uses cases for those who want to write the code for their data integration jobs and take advantage of the power that goes with that approach. Databricks provides an ideal platform on which to implement data integration pipelines geared towards machine learning applications.

Data catalog and governance

Azure Data Catalog provides baseline functionality in the governance arena today.

Finally, but not at all less critical, is the subject of data curation and governance. Technologies in this area of analytics help you in two ways: they can catalog everything across your data landscape, making data in data warehouses, and especially in data lakes (which are sometimes difficult to navigate), far more discoverable. Data governance tools can also help you manage and protect your data so that more sensitive data is only visible to the people with appropriate permissions.

Azure Data Catalog provides baseline functionality in the governance arena today. It tracks the metadata of your data sets and makes it searchable, enabling members of your team to find the data they'll need for their work, more easily. Azure Data Catalog also lets you "tag" your data sets and columns within them—essentially assigning keywords to them, to make them more searchable still.

Through the HDInsight Application Platform, third-party products, like Waterline Data and Unifi, are available to help with further governance needs, including automated identification of personally identifiable information (PII), data classification, and machine learning-enabled automation of tagging. Here we see again how the Azure cloud platform takes care of core capabilities and many advanced ones, then integrates ecosystem solutions at the top of the stack for fit, finish, and customized solutions.

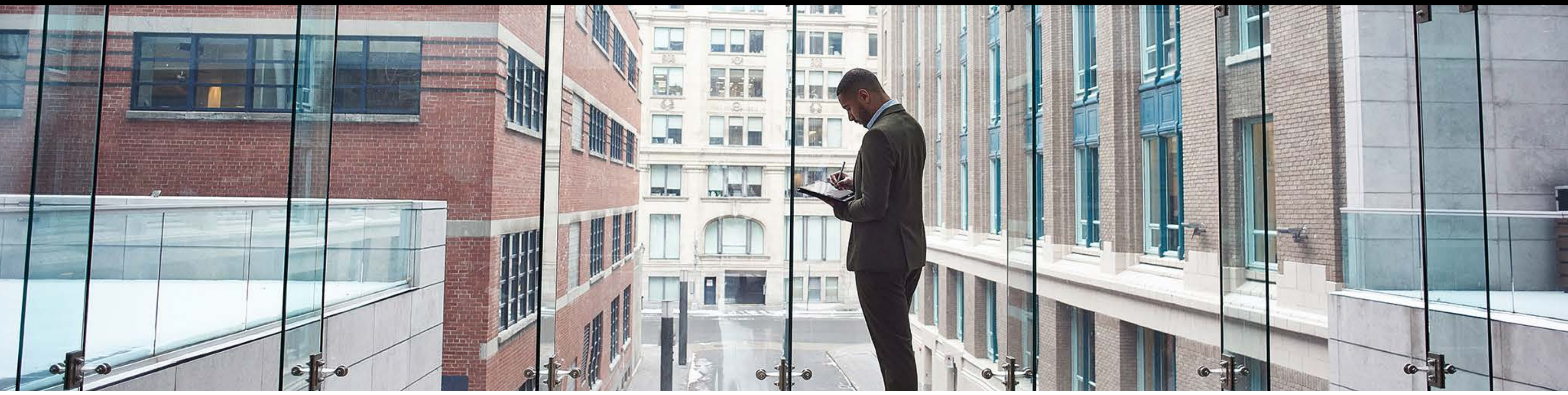
What it means...

A comprehensive understanding of these building blocks will support the right implementation for the organization and help harness the full power of cloud scale analytics, bringing modern data warehouse and real time analytics to life.

Chapter 6

How AI fits in

Thus far we've seen the combinative power of cloud scale analytics building blocks. The data warehouse, data lake, BI, data virtualization, storage technologies, integration, and governance in the cloud are complimentary. They bring the notion of the modern data warehouse and real-time analytics past rhetoric and vision to realistic implementation. But what about machine learning (ML) and artificial intelligence (AI)?



Everything in a well-crafted analytics initiative will lay solid foundation for follow-on AI work, making the tie-ins quite strong.

The short answer is that everything in a well-crafted analytics initiative will lay solid foundation for follow-on AI work, making the tie-ins quite strong indeed.

Azure Machine Learning. Azure Machine Learning provides a full ML platform, suitable for experimentation, training, production deployment and scoring against deployed ML models at scale. Azure ML is available in virtually any environment where developers may prefer to work with the Python programming language, which includes Visual Studio Code, PyCharm and notebooks on Azure Databricks or HDInsight.

Azure Databricks. Azure Databricks itself offers an ML platform, called Spark MLlib, and a new ML experimentation facility called MLFlow. Developers can mix and match data engineering, streaming, analytics, and ML into single notebooks, or jobs. And, as we've mentioned, Azure ML can be integrated into the Databricks environment.

Microsoft Machine Learning Server

Want more? Microsoft Machine Learning Server, available in the cloud on HDInsight, on-premises on a standalone basis and in-database with both Azure SQL Database and SQL Server, provides another avenue. Developers can work with the R language's built-in ML facilities, and, in the case of SQL Server, various Python packages as well, including Scikit-learn, PyTorch, and even deep learning framework TensorFlow. Models developed in many of these environments can be moved around to others, providing a comprehensive array of choices.

Models developed in many of these environments can be moved around to others, providing a comprehensive array of choices.

Microsoft's Automated Machine Learning. If you want a more straightforward route to ML, check out Microsoft's Automated Machine Learning (AutoML) technology. It requires less data science background and can cost less. Available as part of Azure Machine Learning and integrated with Power BI Premium, Microsoft AutoML eliminates much of the mystery involved in picking the right algorithms and tuning their hyperparameters. It also opens AI to the broader developer world, through its integration with ML.NET.

Azure Cognitive Services. Developers and others may also be very interested in Azure Cognitive Services. These are ready-to-run serverless cloud services built around pre-trained models. Developers need only bring their data sets to these services and call them to run sophisticated predictions in many realms, including vision and natural language. Cognitive services are also integrated into Power BI Premium, as are models deployed to Azure Machine Learning.

What it means...

The opportunity for developers to apply Microsoft's Automated Machine Learning and Azure Cognitive Services to modern data warehouses along with real time analytics will enable the next level of cloud scale analytics with sophisticated predictions including vision and natural language.

Summary

We've looked at the what and why of cloud-scale analytics, and we've seen how analytics service building blocks – including data warehouse, data lake, BI, data virtualization, storage, integration, governance and AI – work together to make it a reality.

Microsoft offers a variety of cloud scale analytics services that leverage the best of open source and Microsoft's unique data platform technology. Sometimes, such a broad array of options can be a little intimidating; the best antidote is to get hands-on with one or two of the technologies, then branch out further. Microsoft Azure offers a free tier that lets customers do just that; Power BI Desktop is available as a free download and the Power BI service offers a free subscription level.

There's no magic involved in analytics. It's simply a matter of emancipating the information.

How to get started

Microsoft also offers a variety of learning resources, both formal and informal, that can make the hands-on experience more productive and fun. And for those who wish to explore Microsoft's cloud scale analytics services further before getting hands-on, a number of additional overview resources are available. A partial list of these learning and overview resources appears in Appendix A.

Envision your organization in full mastery of modern data warehouse, real-time analytics and even machine learning.

When your learning work is complete, you'll want to move into implementing production solutions. But don't just dive into it; instead, plan ahead to get the best return on your investment. Identify your data sources. Brainstorm several use cases – including feasibly short-term wins, and goals for longer-term innovative achievements. Look for operational challenges your business is having that cloud scale analytics could help you address. Identify where your on-premises systems have helped, but have also hit barriers – in terms of expandability, workload flexibility and resource elasticity – and which cloud scale analytics technologies and services could help you transcend those barriers.

Think about areas of your operations or business where you'd like to be more proactive, where you want to understand better what happened and where you could benefit from being able to predict what will happen. Set the milestones for your journey, bring in trusted partners to help you set out on it, and identify prudent goals so you can measure and monitor your success, and course correct where necessary.

Envision your organization in full mastery of modern data warehouse, real-time analytics and even machine learning. That's your cloud scale analytics destination. That's your motivation. That's your incentive.

Then recall what we said toward the beginning of this e-book: that there's no magic involved in analytics. It's simply a matter of emancipating the information that lies dormant in your data and – in a thoughtful, carefully planned way – using the building block services and technologies that can make that discovery prudent, feasible, repeatable and operationalized.

Get Started Today

Learning Resources

Get started with 12 months of free services

Connect with an Azure sales specialist on pricing, analytics best practices, setting up a proof of concept, and more

Learn why customers are choosing Azure for their analytics