Why should I care about big data?

patterns & practices
proven practices for predictable results

From: Developing big data solutions on Microsoft Azure HDInsight

Many people consider big data solutions to be a new way to do data warehousing when the volume of data exceeds the capacity or cost limitations for relational database systems. However, it can be difficult to fully grasp what big data solutions really involve, what hardware and software they use, and how and when they are useful. There are some basic questions, the answers to which will help you understand where big data solutions are useful—and how you might approach the topic in terms of implementing your own solutions:

Why do I need a big data solution?

In the most simplistic terms, organizations need a big data solution to enable them to survive in a rapidly expanding and increasingly competitive market where the sources and the requirements to store data are growing at exponential rate. Big data solutions are typically used for:

- **Storing huge volumes of unstructured or semi-structured data**. Many organizations need to handle vast quantities of data as part of their daily operations. Examples include financial and auditing data, and medical data such as patients' notes. Processing, backing up, and querying all of this data becomes more complex and time consuming as the volume increases. Big data solutions are designed to store vast quantities of data (typically on distributed servers with automatic generation of replicas to guard against data loss), together with mechanisms for performing queries on the data to extract the information the organization requires.
- **Finding hidden insights in large stores of data**. For example, organizations want to know how their products and services are perceived in the market, what customers think of the organization, whether advertising campaigns are working, and which facets of the organization are (or are not) achieving their aims. Organizations typically collect data that is useful for generating business intelligence (BI) reports, and to provide input for management decisions. However, they are increasingly implementing mechanisms that collect other types of data such as “sentiment data” (emails, comments from web site feedback mechanisms, and tweets that are related to the organization's products and services), click-through data, information from sensors in users' devices (such as location data), and website log files.
- **Extracting vital management information**. The vast repositories of data often contain useful, and even vital information that can be used for product and service planning, coordinating advertising campaigns, improving customer service, or as an input to reporting systems. This information is also very useful for predictive analysis such as estimating future profitability in a financial scenario, or for an insurance company to predict the possibility of accidents and claims. Big data solutions allow you to store and extract all this information, even if you don’t know when or how you will use the data at the time you are collecting it.

Successful organizations typically measure performance by discovering the customer value that each part of their

operation generates. Big data solutions provide a way to help you discover value, which often cannot be measured just through traditional business methods such as cost and revenue analysis.

What problems do big data solutions solve?

Big data solutions were initially seen to be primarily a way to resolve the limitation with traditional database systems due to:

- **Volume**: Big data solutions are designed and built to store and process hundreds of terabytes, or even petabytes of data in a way that can dramatically reduce storage cost, while still being able to generate BI and comprehensive reports.
- **Variety**: Organizations often collect unstructured data, which is not in a format that suits relational database systems. Some data, such as web server logs and responses to questionnaires may be preprocessed into the traditional row and column format. However, data such as emails, tweets, and web site comments or feedback, are semi-structured or even unstructured data. Deciding how to store this data using traditional database systems is problematic, and may result in loss of useful information if the data must be constricted to a specific schema when it is stored.

> **Note:**
>
> Big data solutions typically target scenarios where there is a huge volume of unstructured or semi-structured data that must be stored and queried to extract business intelligence. Typically, the majority of data currently stored in big data solutions is unstructured or semi-structured.

- **Velocity**: The rate at which data arrives may make storage in an enterprise data warehouse problematic, especially where formal preparation processes such as examining, conforming, cleansing, and transforming the data must be accomplished before it is loaded into the data warehouse tables.

The combination of all these factors means that, in some circumstances, a big data batch processing solution may be a more practical proposition than a traditional relational database system. However, as big data solutions have continued to evolve it has become clear that they can also be used in a fundamentally different context: to quickly get insights into data, and to provide a platform for further investigation in a way that just isn't possible with traditional data storage, management, and querying tools.

Figure 1 demonstrates how you might go from a semi-intuitive guess at the kind of information that might be hidden in the data, to a process that incorporates that information into your business domain.

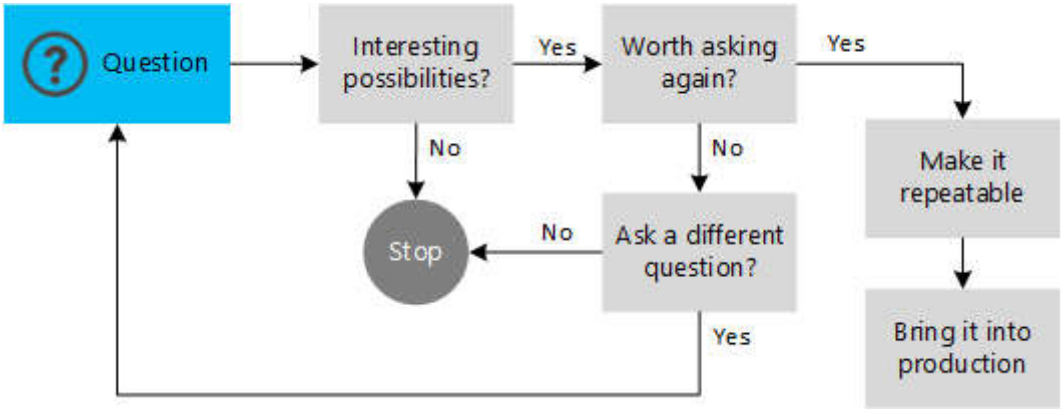

Figure 1 - Big data solutions as an experimental data investigation platform

As an example, you may want to explore the postings by users of a social website to discover what they are saying about your company and its products or services. Using a traditional BI system would mean waiting for the database architect and administrator to update the schemas and models, cleanse and import the data, and design suitable reports. But it's unlikely that you'll know beforehand if the data is actually capable of providing any useful information, or how you might go about discovering it. By using a big data solution you can investigate the data by asking any questions that may seem relevant. If you

find one or more that provide the information you need you can refine the queries, automate the process, and incorporate it into your existing BI systems.

> Big data solutions aren't all about business topics such as customer sentiment or web server log file analysis. They have many diverse uses around the world and across all types of applications. Police forces are using big data techniques to predict crime patterns, researchers are using them to explore the human genome, particle physicists are using them to search for information about the structure of matter, and astronomers are using them to plot the entire universe. Perhaps the last of these really is a big "big data" solution!

How is a big data solution different from traditional database systems?

Traditional database systems typically use a relational model where all the data is stored using predetermined schemas, and linked using the values in specific columns of each table. Requiring a schema to be applied when data is written may mean that some information hidden in the data is lost. There are some more flexible mechanisms, such as the ability to store XML documents and binary data, but the capabilities for handling these types of data are usually quite limited.

Big data solutions do not force a schema onto the stored data. Instead, you can store almost any type of structured, semi-structured, or unstructured data and then apply a suitable schema when you query this data. Big data solutions store the data in its raw format and apply a schema only when the data is read, which preserves all of the information within the data.

Traditional database systems typically consist of a central node where all processing takes place, which means that all the data from storage must be moved to the central location for processing. The capacity of this central node can be increased only by scaling up, and there is a physical limitation on the number of CPUs and memory, depending on the chosen hardware platform. The consequence of this is a limitation of processing capacity, as well as network latency when the data is moved to the central node.

In contrast, big data solutions are optimized for storing vast quantities of data using simple file formats and highly distributed storage mechanisms, and the initial processing of the data occurs at each storage node. This means that, assuming you have already loaded the data into the cluster storage, the bulk of the data does not need to be moved over the network for processing.

Figure 2 shows some of the basic differences between a relational database system and a big data solution in terms of storing and querying data. Notice how both relational databases and big data solutions use a cluster of servers; the main difference is where query processing takes place and how the data is moved across the network.

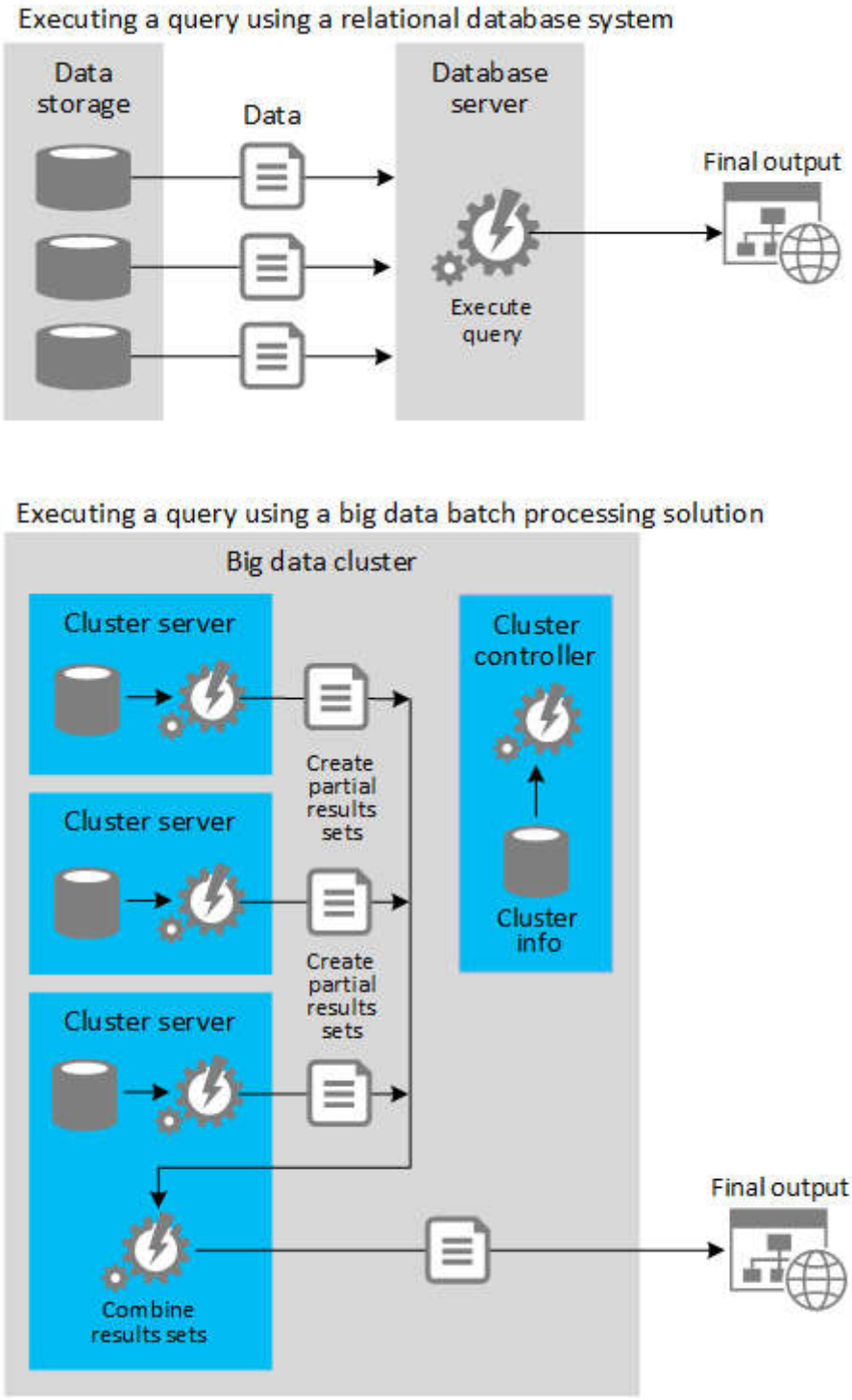

Figure 2 - Some differences between relational databases and big data batch processing solutions

Modern data warehouse systems typically use high speed fiber networks, in-memory caching, and indexes to minimize data transfer delays. However, in a big data solution only the results of the distributed query processing are passed across the cluster network to the node that will assemble them into a final results set. Under ideal conditions, performance during the initial stages of the query is limited only by the speed and capacity of connectivity to the co-located disk subsystem, and this initial processing occurs in parallel across all of the cluster nodes.

Note:
The servers in a cluster are typically co-located in the same datacenter and connected over a low-latency, high-bandwidth

network. However, big data solutions can work well even without a high capacity network, and the servers can be more widely distributed, because the volume of data moved over the network is much less than in a traditional relational database cluster.

The ability to work with highly distributed data and simple file formats also opens up opportunities for more efficient and more comprehensive data collection. For example, services and applications can store data in any of the predefined distributed locations without needing to preprocess it or execute queries that can absorb processing capacity. Data is simply appended to the files in the data store. Any processing required on the data is done when it is queried, without affecting the original data and risking losing valuable information.

Queries to extract information in a big data solution are typically batch operations that, depending on the data volume and query complexity, may take some time to return a final result. However, when you consider the volumes of data that big data solutions can handle, the fact that queries run as multiple tasks on distributed servers does offer a level of performance that may not be achievable by other methods. While it is possible to perform real-time queries, typically you will run the query and store the results for use within your existing BI tools and analytics systems. This means that, unlike most SQL queries used with relational databases, big data queries are typically not executed repeatedly as part of an application's execution—and so batch operation is not a major disadvantage.

Big data systems are also designed to be highly resilient against failure of storage, networks, and processing. The distributed processing and replicated storage model is fault-tolerant, and allows easy re-execution of individual stages of the process. The capability for easy scaling of resources also helps to resolve operational and performance issues.

The following table summarizes the major differences between a big data solution and existing relational database systems.

Feature	**Relational database systems**	**Big data solutions**
Data types and formats	Structured	Semi-structured and unstructured
Data integrity	High—transactional updates	Depends on the technology used—often follows an eventually consistent model
Schema	Static—required on write	Dynamic—optional on read and write
Read and write pattern	Fully repeatable read/write	Write once, repeatable read
Storage volume	Gigabytes to terabytes	Terabytes, petabytes, and beyond
Scalability	Scale up with more powerful hardware	Scale out with additional servers
Data processing distribution	Limited or none	Distributed across the cluster
Economics	Expensive hardware and software	Commodity hardware and open source software

Will a big data solution replace my relational databases?

Big data batch processing solutions offer a way to avoid storage limitations, or to reduce the cost of storage and processing, for huge and growing volumes of data; especially where this data might not be part of a vital business function. But this isn't to say that relational databases have had their day. Continual development of the hardware and software for this core business function provides capabilities for storing very large amounts of data. For example, Microsoft Analytics Platform System (APS) can store hundreds of terabytes of data.

In fact, the relational database systems we use today and the more recent big data batch processing solutions are complementary mechanisms. Big data batch processing solutions are extremely unlikely ever to replace the existing relational database—in the majority of cases they complement and augment the capabilities for managing data and generating BI. For example, it's common to use a big data query to create a result set that is then stored in a relational database for use in the generation of BI, or as input to another process.

Big data is also a valuable tool when you need to handle data that is arriving very quickly, and which you can process later. You can dump the data into the storage cluster in its original format, and then process it when required using a query that extracts the required result set and stores it in a relational database, or makes it available for reporting. Figure 3 shows this approach in schematic form.

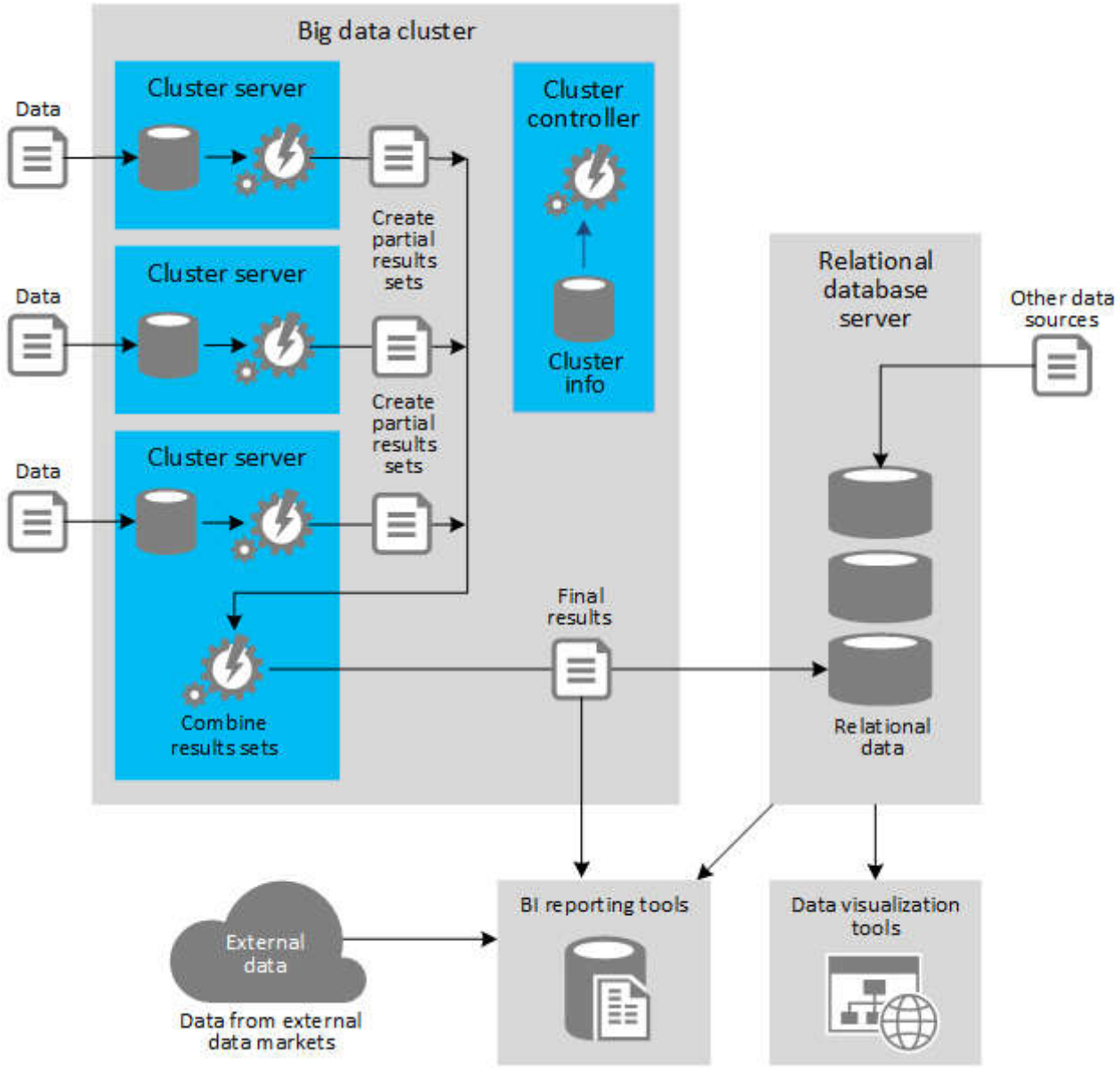

Figure 3 - Combining big data batch processing with a relational database

In this kind of environment, additional capabilities are enabled. Big data batch processing systems work with almost any type of data. It is quite feasible to implement a bidirectional data management solution where data held in a relational database or BI system can be processed by the big data batch processing mechanism, and fed back into the relational database or used for analysis and reporting. This is exactly the type of environment that Microsoft Analytics Platform System (APS) provides.

Note:
The topic Use case 4: BI integration in this guide explores and demonstrates the integration of a big data solution with existing data analysis tools and enterprise BI systems. It describes the three main stages of integration, and explains the benefits of each approach in terms of maximizing the usefulness of your big data implementations.