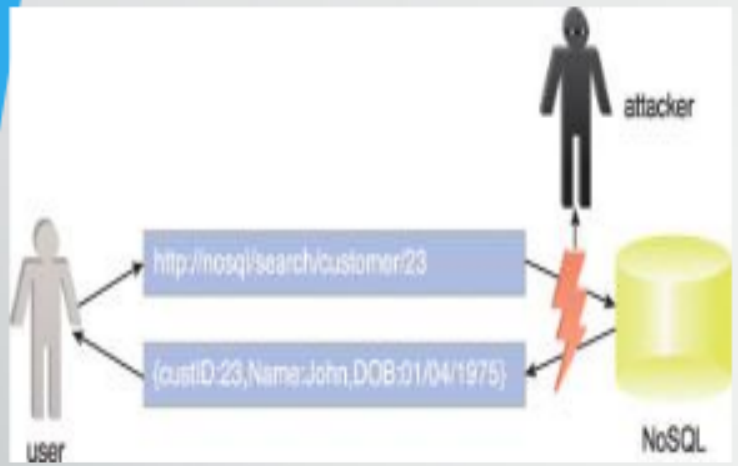
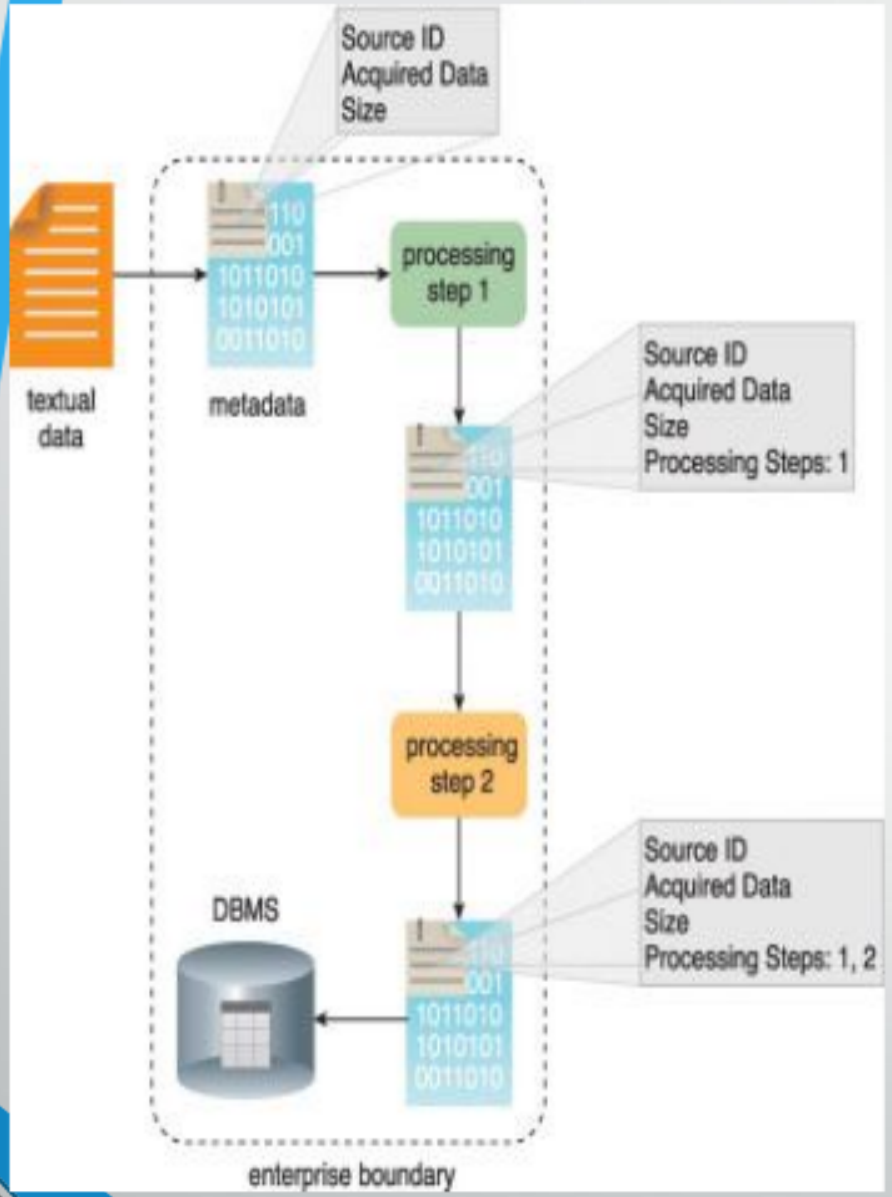
Security



- Some of the components of Big Data solutions lack the robustness of traditional enterprise solution environments when it comes to access control and data security.
- Securing Big Data involves ensuring that the data networks and repositories are sufficiently secured via authentication and authorization mechanisms.
- Big Data security further involves establishing data access levels for different categories of users.
- For example, unlike traditional relational database management systems, NoSQLdatabases generally do not provide robust built-in security mechanisms.
- They instead rely on simple HTTP-based APIs where data is exchanged in plaintext, making the data prone to network-based attacks,

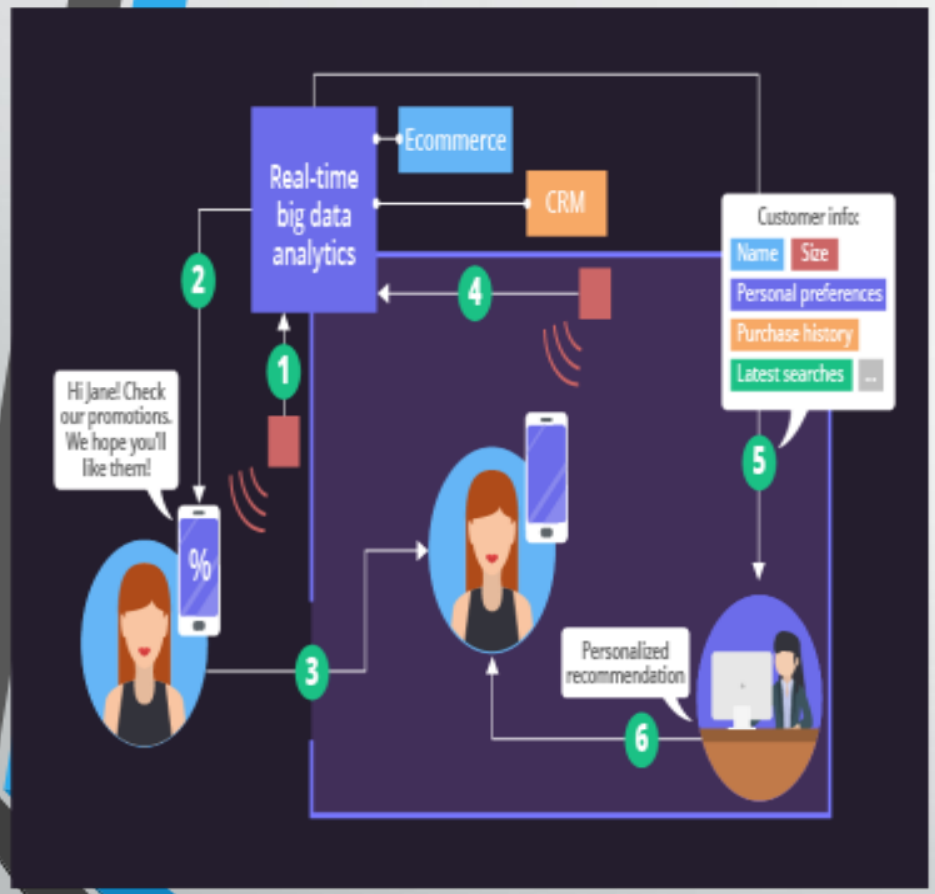
Provenance

- Provenance refers to information about the source of the data and how it has been processed.
- Provenance information helps determine the authenticity and quality of data, and it can be used for auditing purposes.
- Maintaining provenance as large volumes of data are acquired, combined and put through multiple processing stages can be a complex task.
- At different stages in the analytics lifecycle, data will be in different states due to the fact it may be being transmitted, processed or in storage.
- These states correspond to the notion of data-in-motion, data-in-use and data-at-rest.
- Importantly, whenever Big Data changes state, it should trigger the capture of provenance information that is recorded as metadata.



- As data enters the analytic environment, its provenance record can be initialized with the recording of information that captures the pedigree of the data.
- Ultimately, the goal of capturing provenance is to be able to reason over the generated analytic results with the knowledge of the origin of the data and what steps or algorithms were used to process the data that led to the result.
- Provenance information is essential to being able to realize the value of the analytic result.
- Much like scientific research, if results cannot be justified and repeated, they lack credibility.
- When provenance information is captured on the way to generating analytic results the results can be more easily trusted and thereby used with confidence.

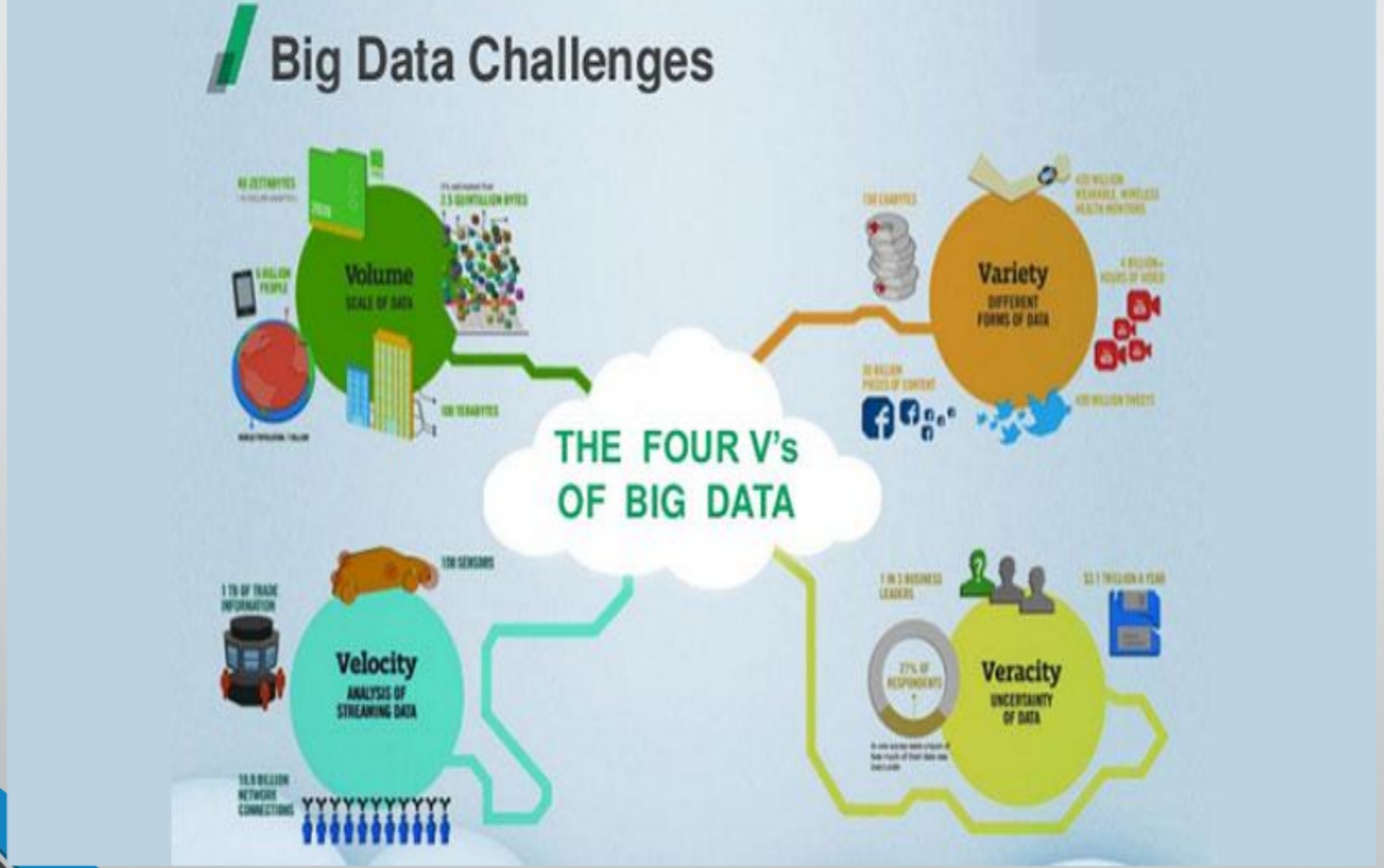
Limited Real time Support



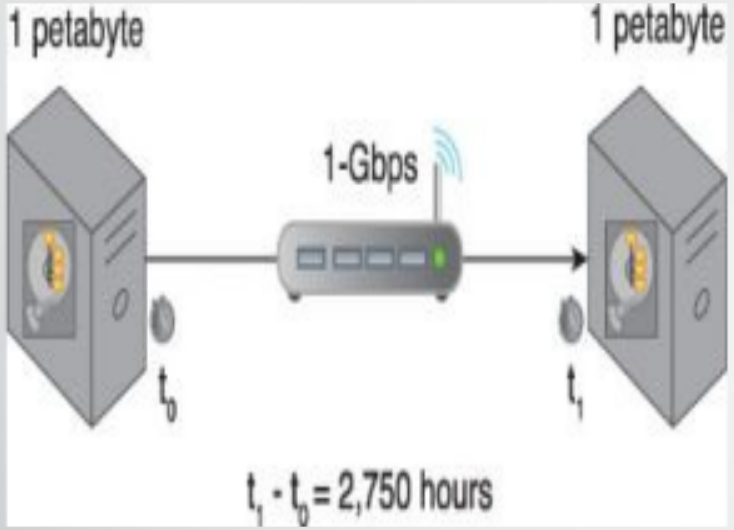
- Dashboards and other applications that require streaming data and alerts often demand real time or near-real time data transmissions.
- Many open source Big Data solutions and tools are batch-oriented; however, there is a new generation of real time capable open source tools that have support for streaming data analysis.
- Many of the real time data analysis solutions that do exist are proprietary.
- Approaches that achieve near-real time results often process transactional data as it arrives and combine it with previously summarized batch-processed data.


XENONSTACK
Big Data
Challenges

Big Data Challenges



Distinct Performance Challenges



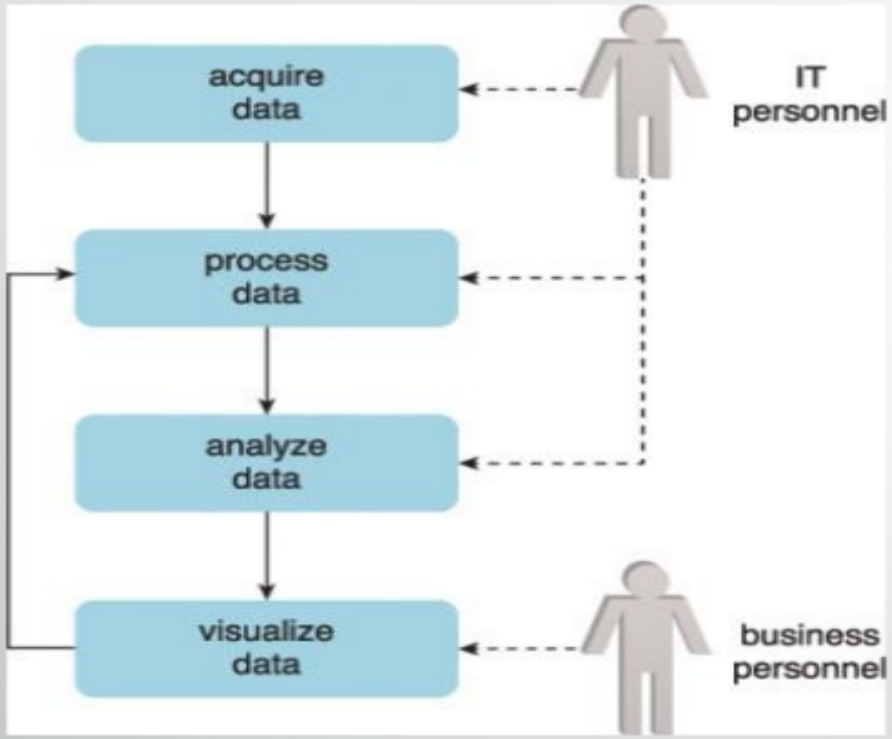
Transferring 1 PB of data via a 1-Gigabit LAN connection at 80% throughput will take approximately 2,750 hours.

- Due to the volumes of data that some Big Data solutions are required to process, performance is often a concern.
- For example, large datasets coupled with complex search algorithms can lead to long query times.
- Another performance challenge is related to network bandwidth.
- With increasing data volumes, the time to transfer a unit of data can exceed its actual data processing time

Distinct Governance Requirements

- Big Data solutions access data and generate data, all of which become assets of the business. A governance framework is required to ensure that the data and the solution environment itself are regulated, standardized and evolved in a controlled manner.
- Examples of what a Big Data governance framework can encompass include:
  - standardization of how data is tagged and the metadata used for tagging
  - policies that regulate the kind of external data that may be acquired
  - policies regarding the management of data privacy and data anonymization
  - policies for the archiving of data sources and analysis results
  - policies that establish guidelines for data cleansing and filtering

Distinct Methodology



Each repetition can help fine-tune processing steps, algorithms and data models to improve the accuracy of results and deliver greater value to the business.

- A methodology will be required to control how data flows into and out of Big Data solutions.
- It will need to consider how feedback loops can be established to enable the processed data to undergo repeated refinement.
- For example, an iterative approach may be used to enable business personnel to provide IT personnel with feedback on a periodic basis.
- Each feedback cycle provides opportunities for system refinement by modifying data preparation or data analysis steps.

Clouds

- Clouds provide remote environments that can host IT infrastructure for large-scale storage and processing, among other things.
- Regardless of whether an organization is already cloud-enabled, the adoption of a Big Data environment may necessitate that some or all of that environment be hosted within a cloud.
- For example, an enterprise that runs its CRM system in a cloud decides to add a Big Data solution in the same cloud environment in order to run analytics on its CRM data.
- This data can then be shared with its primary Big Data environment that resides within the enterprise boundaries.

BIG DATA ADOPTION CONSIDERATIONS

UNIT III

Common justifications for incorporating a cloud environment in support of a Big Data solution include:

- Inadequate in-house hardware resources
- Upfront capital investment for system procurement is not available
- The project is to be isolated from the rest of the business so that existing business processes are not impacted
- The Big Data initiative is a proof of concept
- Datasets that need to be processed are already cloud resident
- The limits of available computing and storage resources used by an in- house Big Data solution are being reached

How is Big Data Related to Cloud Computing?

- Hence, from the above description, we can see that Cloud enables "As-a-Service" pattern by abstracting the challenges and complexity through a scalable and elastic self-service application. Big data requirement is same where distributed processing of massive data is abstracted from the end-users.

- There are multiple benefits of Big data analysis in Cloud.

· Improved analysis

- With the advancement of Cloud technology, big data analysis has become more improved causing better results. Hence, companies prefer to perform big data analysis in the Cloud. Moreover, Cloud helps to integrate data from numerous sources.

· Simplified Infrastructure

- Big Data analysis is a tremendous strenuous job on infrastructure as the data comes in large volumes with varying speeds, and types which traditional infrastructures usually cannot keep up with. As Cloud computing provides flexible infrastructure, which we can scale according to the needs at the time, it is easy to manage workloads.

- · Lowering the cost
- Both Big data and Cloud technology delivers value to organizations by reducing ownership. The Pay-per-user model of Cloud turns CAPEX into OPEX. On the other hand, Apache cut down the licensing cost of Big data which is supposed to be cost millions to build and buy. Cloud enables customers for big data processing without large-scale big data resources. Hence, both Big Data and Cloud technology are driving the cost down for enterprise purposes and bringing value to the enterprise.

. Security and Privacy

Data security and privacy are two major concerns when dealing with enterprise data.

Moreover, when your application is hosted on a Cloud platform due to its open environment and limited user control security becomes a primary concern.

On the other hand, being an open-source application, Big data solution like Hadoop uses a lot of third-party services and infrastructure.

Hence, nowadays the system integrators bring in Private Cloud Solution that is Elastic and Scalable.

Furthermore, it also leverages Scalable Distributed Processing.

- Besides that Cloud data is stored and processed in a central location commonly known as Cloud storage server.
- Along with it the service provider and the customer signs a service level agreement (SLA) to gain the trust between them.
- If require the provider also leverages required an advanced level of security control.
- This enables the security of big data in Cloud computing covering the following issues:
  - Protecting big data from advanced threats.
  - How Cloud service providers maintain storage and data.

- There are rules associated with service level agreements for protecting
  - data
  - capacity
  - scalability
  - security
  - privacy
  - availability of data storage and data growth

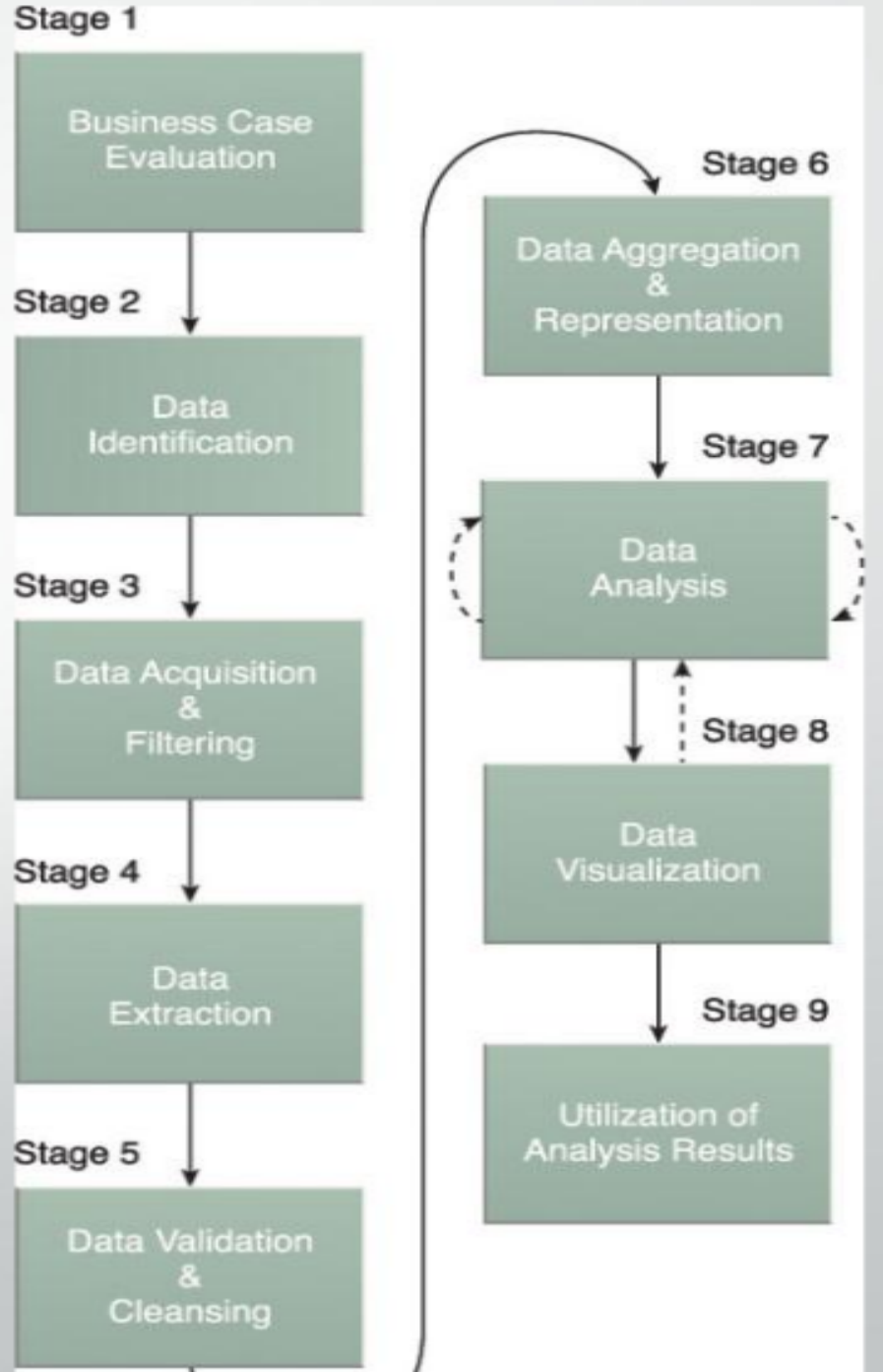
Big Data Analytics Lifecycle

- Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes.
- To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data.

- From a Big Data adoption and planning perspective, it is important that in addition to the lifecycle, consideration be made for issues of training, education, tooling and staffing of a data analytics team.

The Big Data analytics lifecycle can be divided into the following nine stages.

- Business Case Evaluation
- Data Identification
- Data Acquisition & Filtering
- Data Extraction
- Data Validation & Cleansing
- Data Aggregation & Representation
- Data Analysis
- Data Visualization
- Utilization of Analysis Results



TOPICS TO BE COVERED

- Organization Prerequisites
- Data Procurement
- Privacy
- Security
- Provenance
- Limited Real time Support
- Distinct Performance Challenges
- Distinct Governance Requirements
- Distinct Methodology
- Clouds
- Big Data Analytics Lifecycle
- Business Case Evaluation

Virtualization

- Infrastructure plays a crucial role to support any application. Virtualization technology is the ideal platform for big data.
- Virtualized big data applications like Hadoop provide multiple benefits which are not accessible on physical infrastructure, but it simplifies big data Management.

- Big data and Cloud computing point to the convergence of various technologies and trends that makes IT infrastructure and related applications more dynamic, more expendable and more modular
- Hence, Big data and Cloud computing projects rely heavily on virtualization

Organization Prerequisites

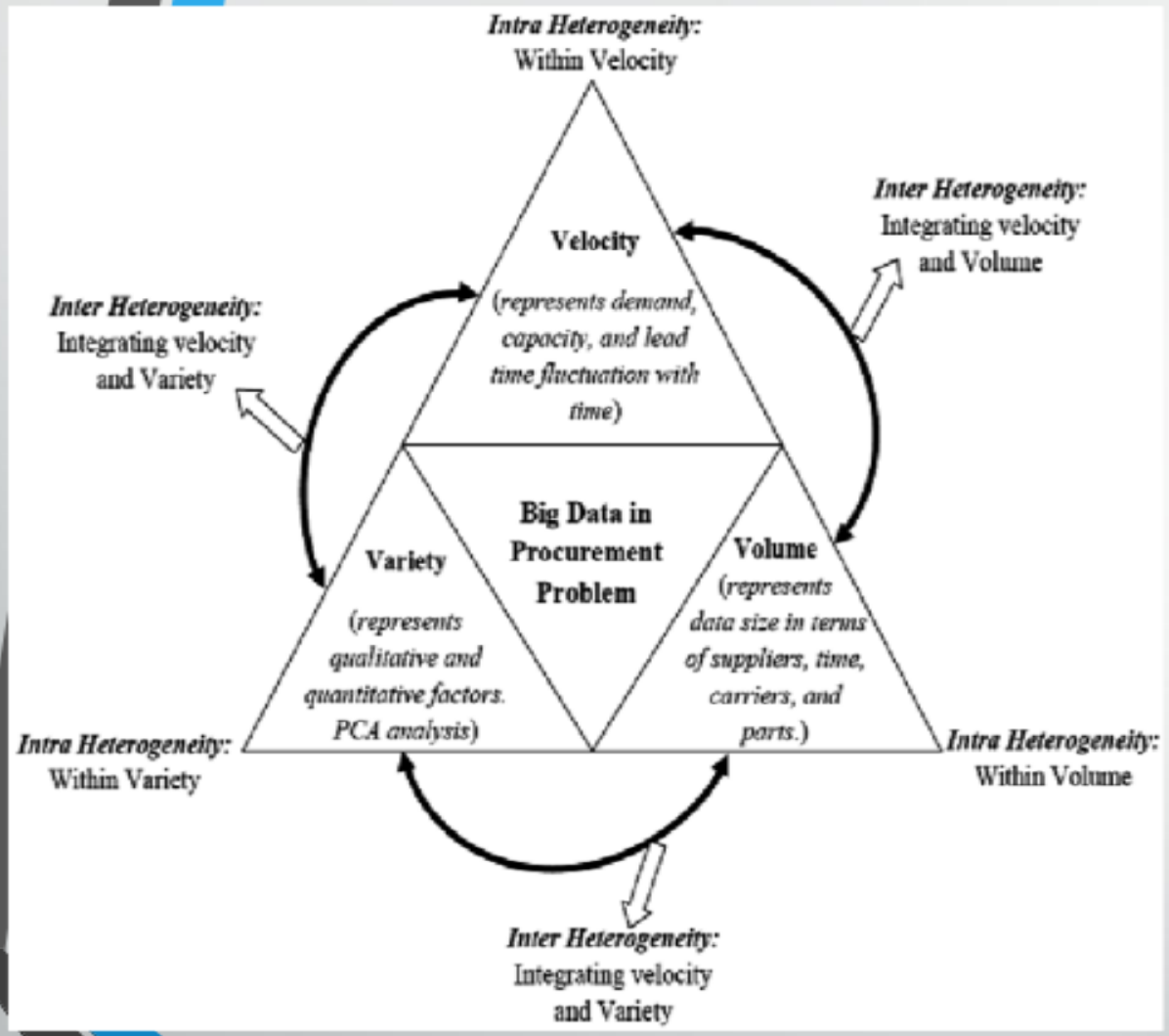
Big Data frameworks are not turn-key solutions.

In order for data analysis and analytics to offer value, enterprises need to have data management and Big Data governance frameworks.

Sound processes and sufficient skillsets for those who will be responsible for implementing, customizing, populating and using Big Data solutions are also necessary.

- Additionally, the quality of the data targeted for processing by Big Data solutions needs to be assessed.
- Outdated, invalid, or poorly identified data will result in low-quality input which, regardless of how good the Big Data solution is, will continue to produce low-quality results.
- The longevity of the Big Data environment also needs to be planned for.
- A roadmap needs to be defined to ensure that any necessary expansion or augmentation of the environment is planned out to stay in sync with the requirements of the enterprise.

Data Procurement

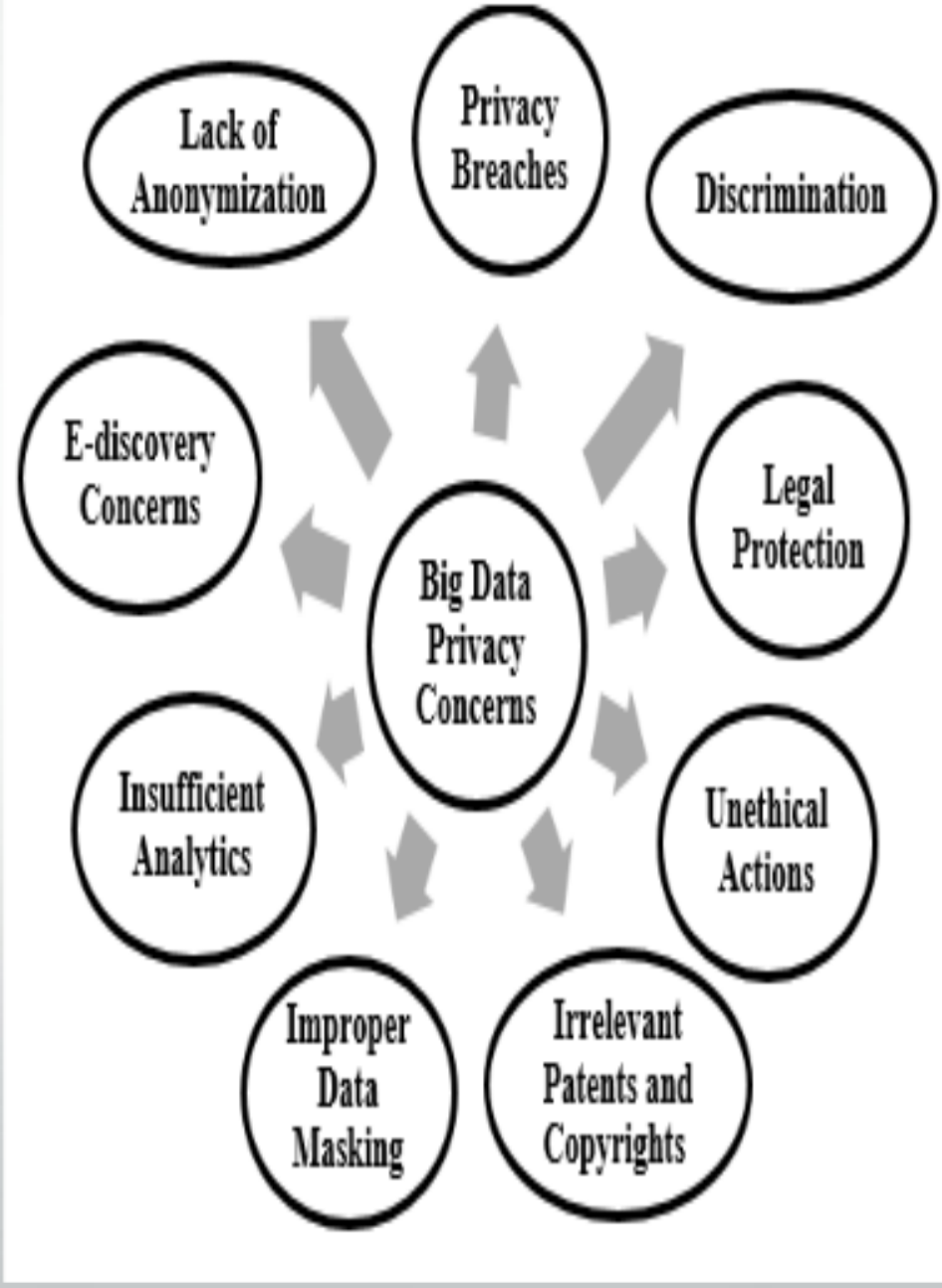


- The acquisition of Big Data solutions themselves can be economical, due to the availability of open-source platforms and tools and opportunities to leverage commodity hardware.
- However, a substantial budget may still be required to obtain external data.
- The nature of the business may make external data very valuable.

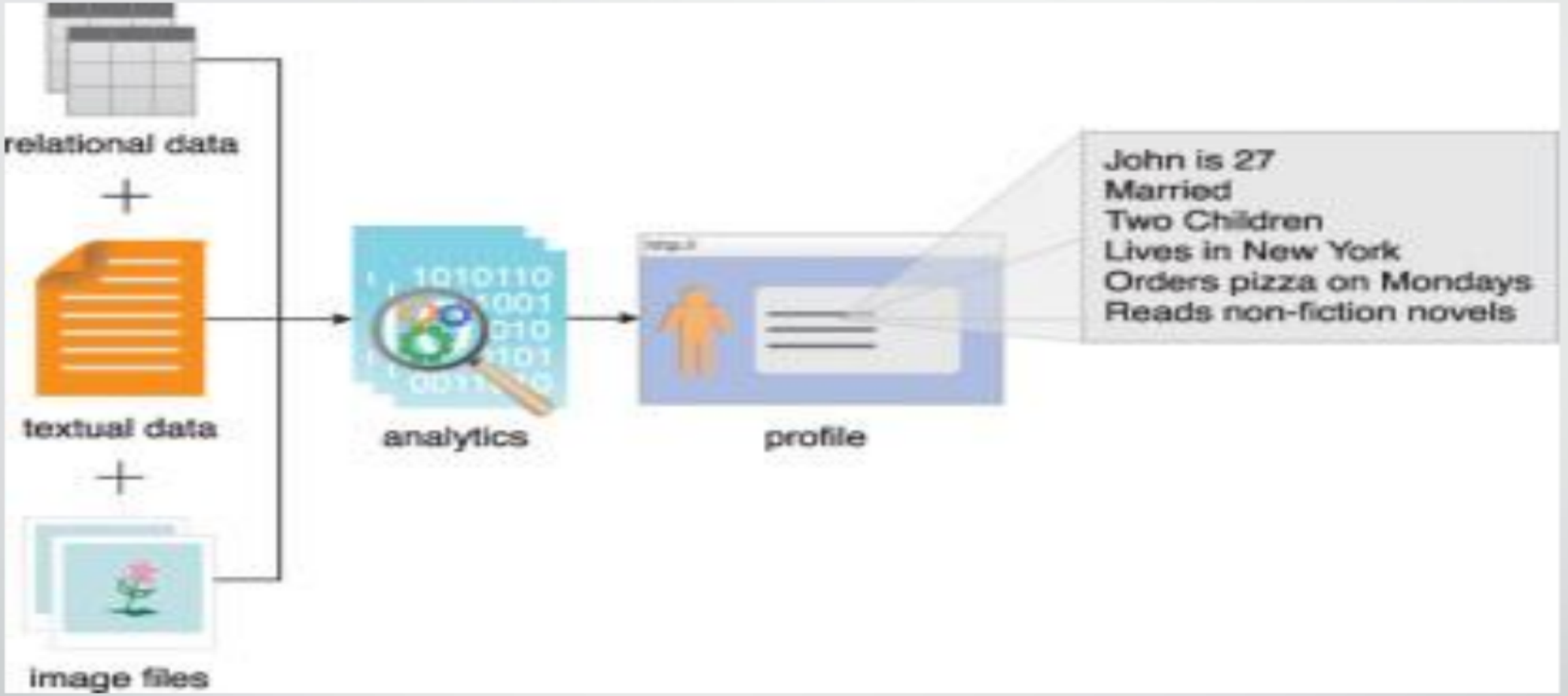
- The greater the volume and variety of data that can be supplied, the higher the chances are of finding hidden insights from patterns.
- External data sources include government data sources and commercial data markets.
- Government-provided data, such as geo-spatial data, may be free.
- However, most commercially relevant data will need to be purchased and may involve the continuation of subscription costs to ensure the delivery of updates to procured datasets

Privacy

- Performing analytics on datasets can reveal confidential information about organizations or individuals.
- Even analyzing separate datasets that contain seemingly benign data can reveal private information when the datasets are analyzed jointly.
- This can lead to intentional or inadvertent breaches of privacy.



- Addressing these privacy concerns requires an understanding of the nature of data being accumulated and relevant data privacy regulations, as well as special techniques for data tagging and anonymization.
- For example, telemetry data, such as a car's GPS log or smart meter data readings, collected over an extended period of time can reveal an individual's location and behavior.


relational data
textual data
image files
analytics
profile
John is 27
Married
Two Children
Lives in New York
Orders pizza on Mondays
Reads non-fiction novels