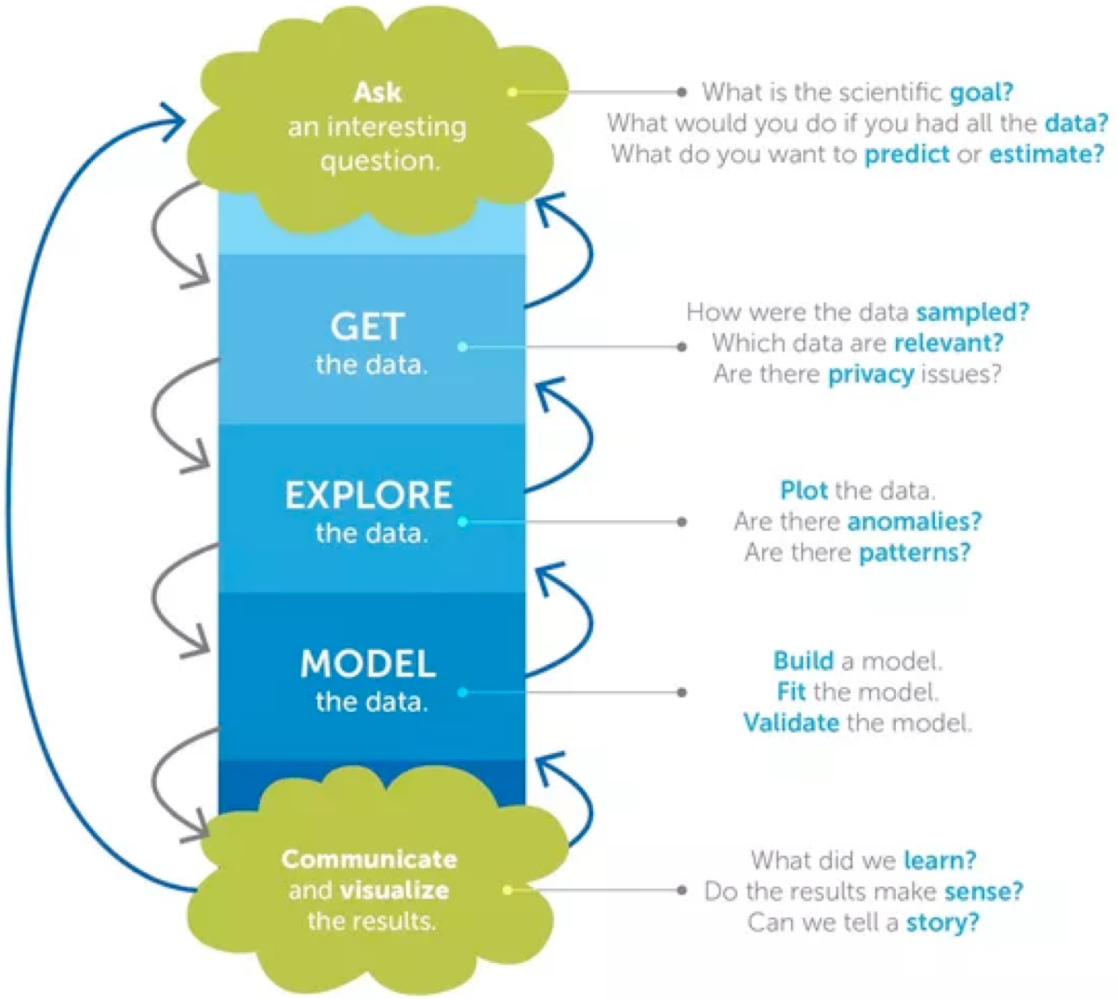
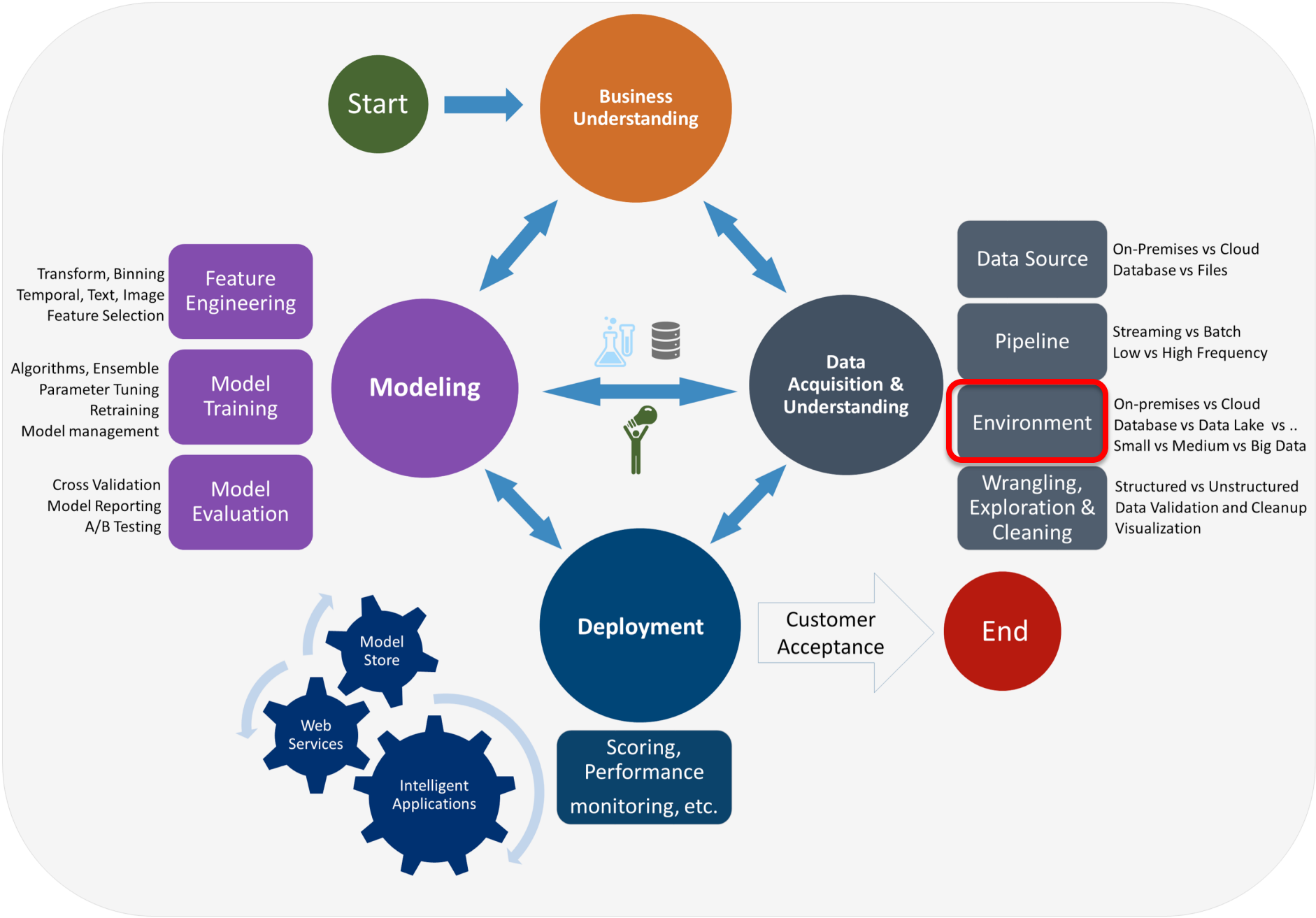
Lecture 2:
Data Science Process and Data Analytics Life Cycle

Benjamin Harvey
The George Washington University
E-mail: bsharve@email.gw.edu

Data Science Process



Data Science Lifecycle



Why Do I Care?

- Data Science is just one aspect of Data Analytics
- Gives insight into what **you need to understand before about the organization and problem space before analytic implementation has started**
- Complete Data Scientist understands the following:
  - Enterprise (Business) Architecture/Systems Engineering
  - Cloud Computing
  - Big Data
  - Open Source Infrastructure
  - Data Science and Analytics
- Business, Applications, Information, Math & Technology (e.g. Research DS vs. Enterprise DS)

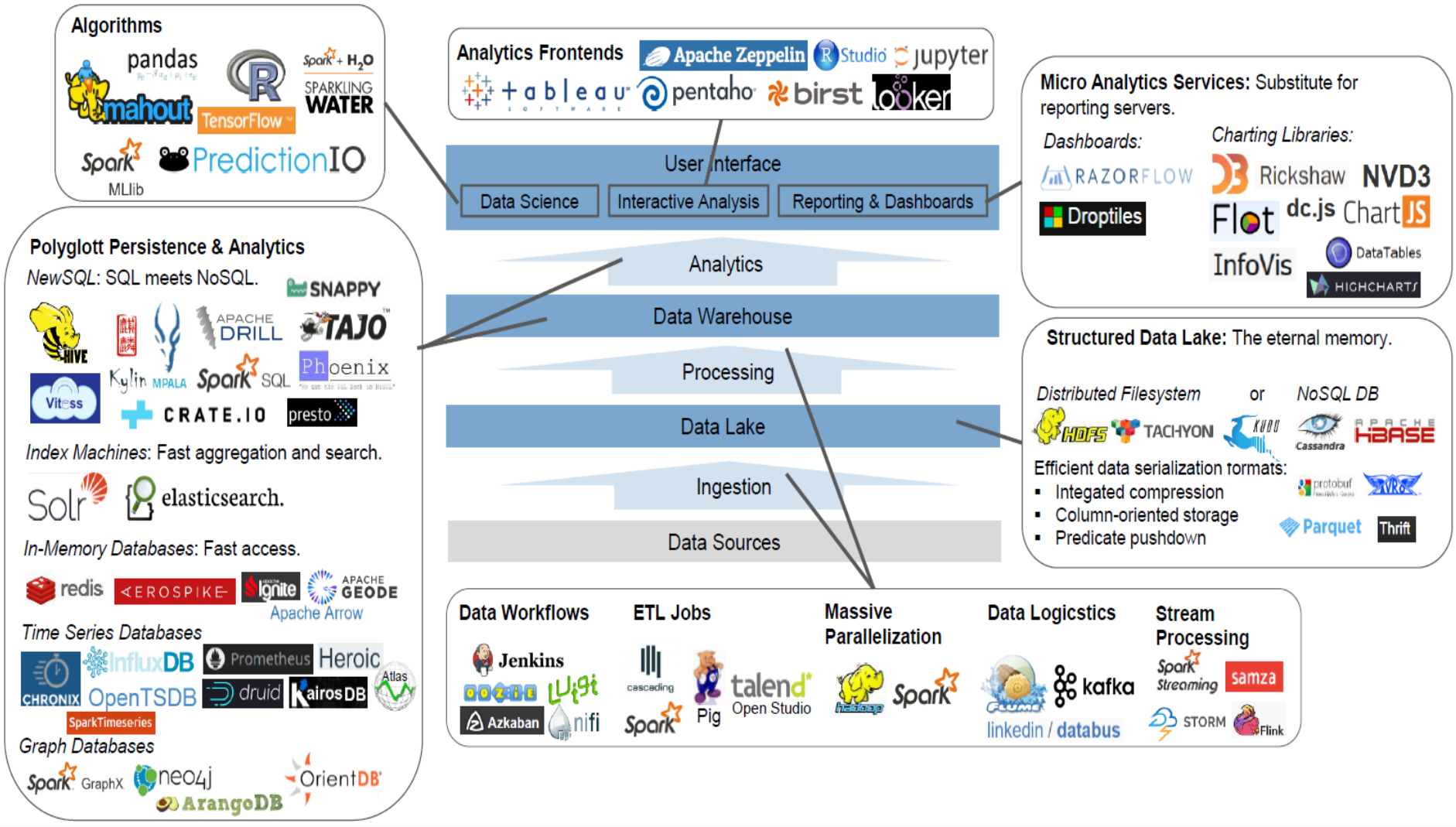
Big Data Planning Considerations

- Organization Prerequisites: Big Data Frameworks
  - Organization Perquisites: Big Data frameworks
- Processing frameworks compute over the data in the system, either by reading from **non-volatile storage or as it is ingested into the system**. Computing over data is the process of extracting information and insight from large quantities of individual data points.
- Types of Big Data frameworks include:
  - **Batch-only frameworks**
  - **Stream-only frameworks**
  - **Hybrid frameworks**

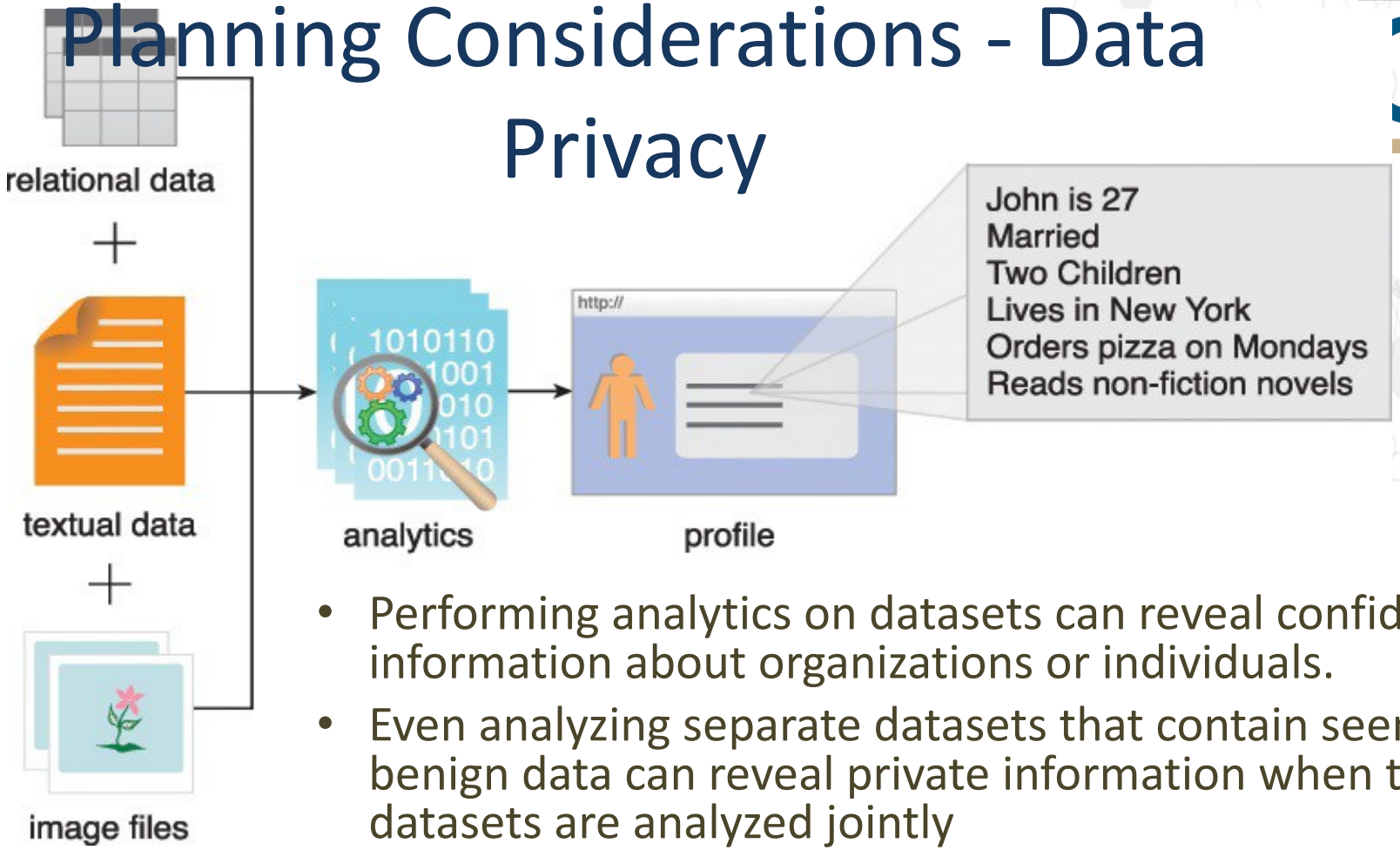
Big Data Planning Considerations

- Technology and Data Procurement
  - The acquisition of Big Data solutions themselves can be economical, due to the availability of free **and open-source (FOSS) platforms and tools** and opportunities to leverage commodity hardware.
  - External data sources include government data sources and commercial data markets.

Data Analytics Life Cycle



Planning Considerations - Data 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
- 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.

Joint Analytics and Privacy

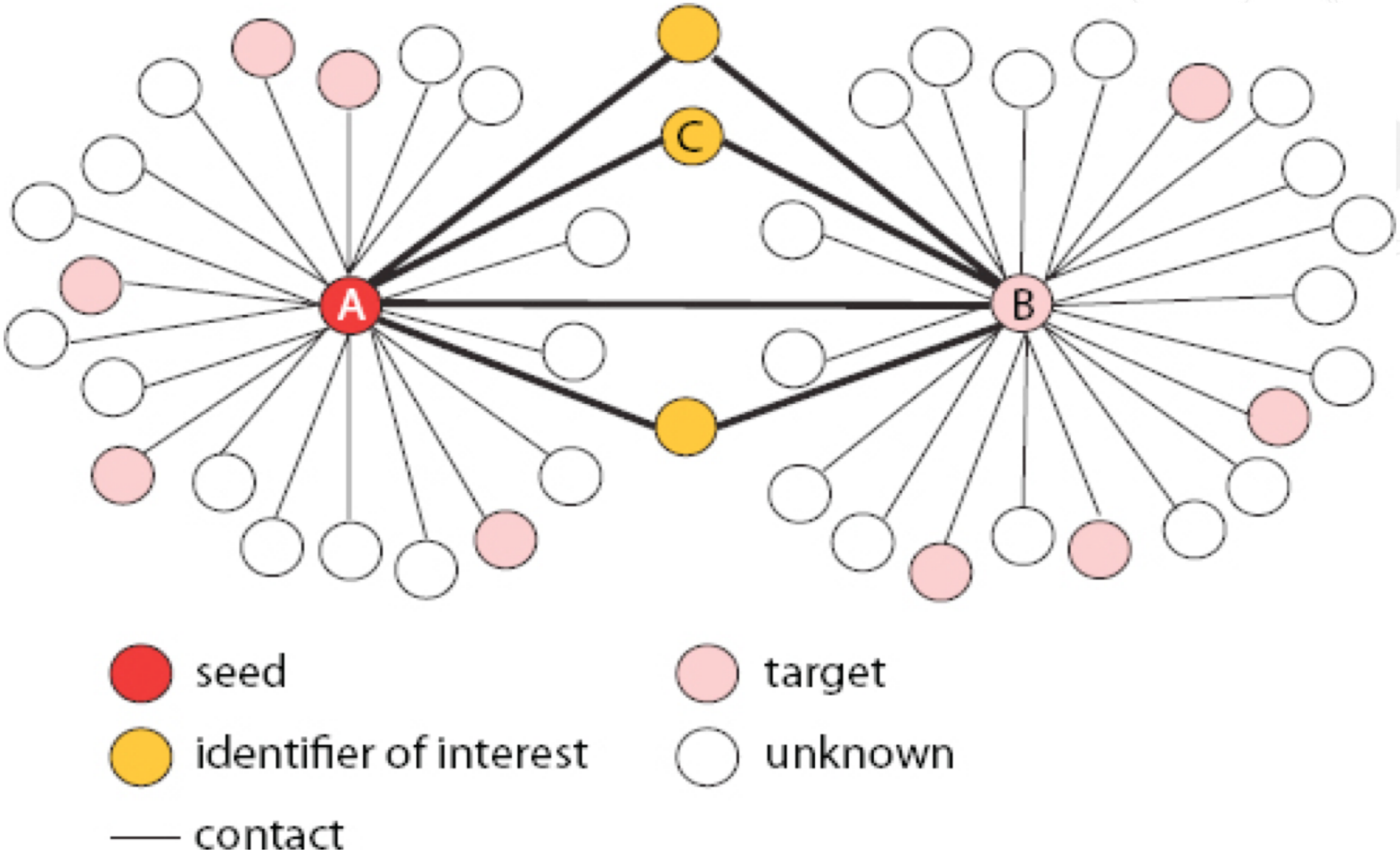


FIGURE 3.1 A network of contacts among identifiers

https://www.nap.edu/read/19414/chapter/5#43

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 (also based upon job role). E.g. SQL injection

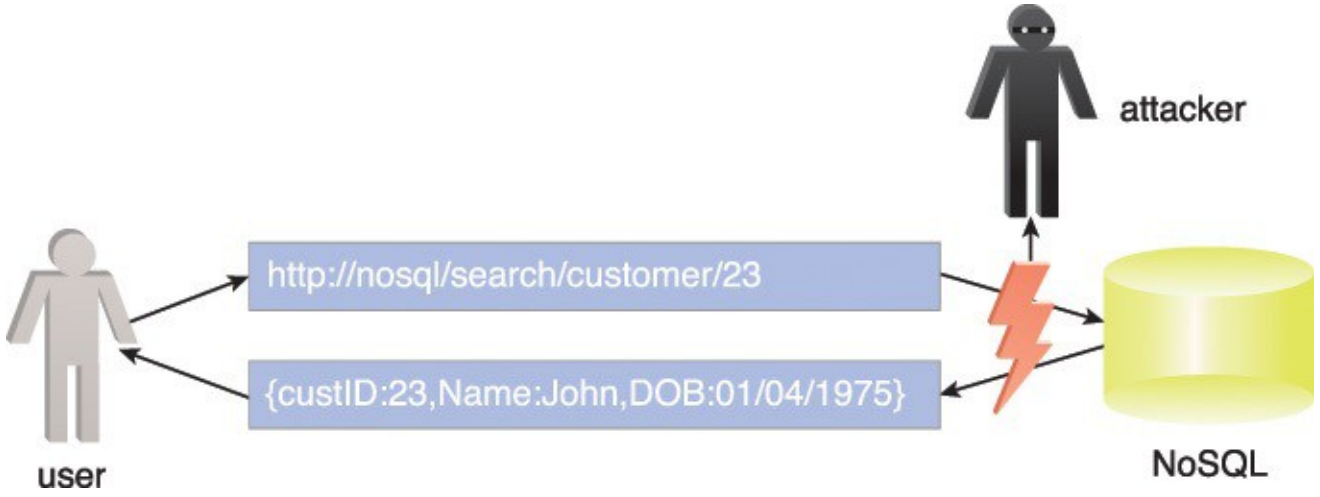
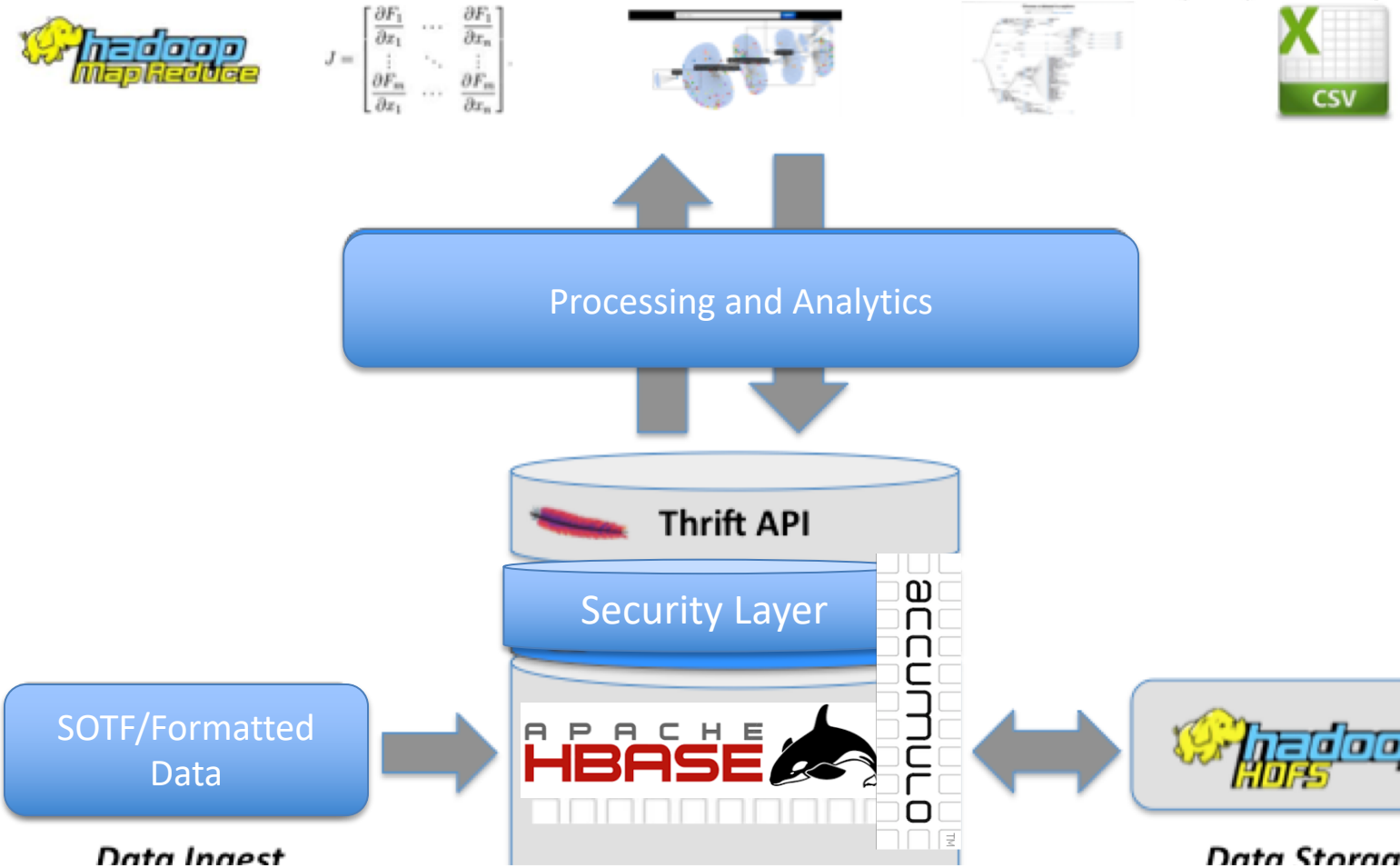


Figure 3.2 NoSQL databases can be susceptible to network-based attacks.

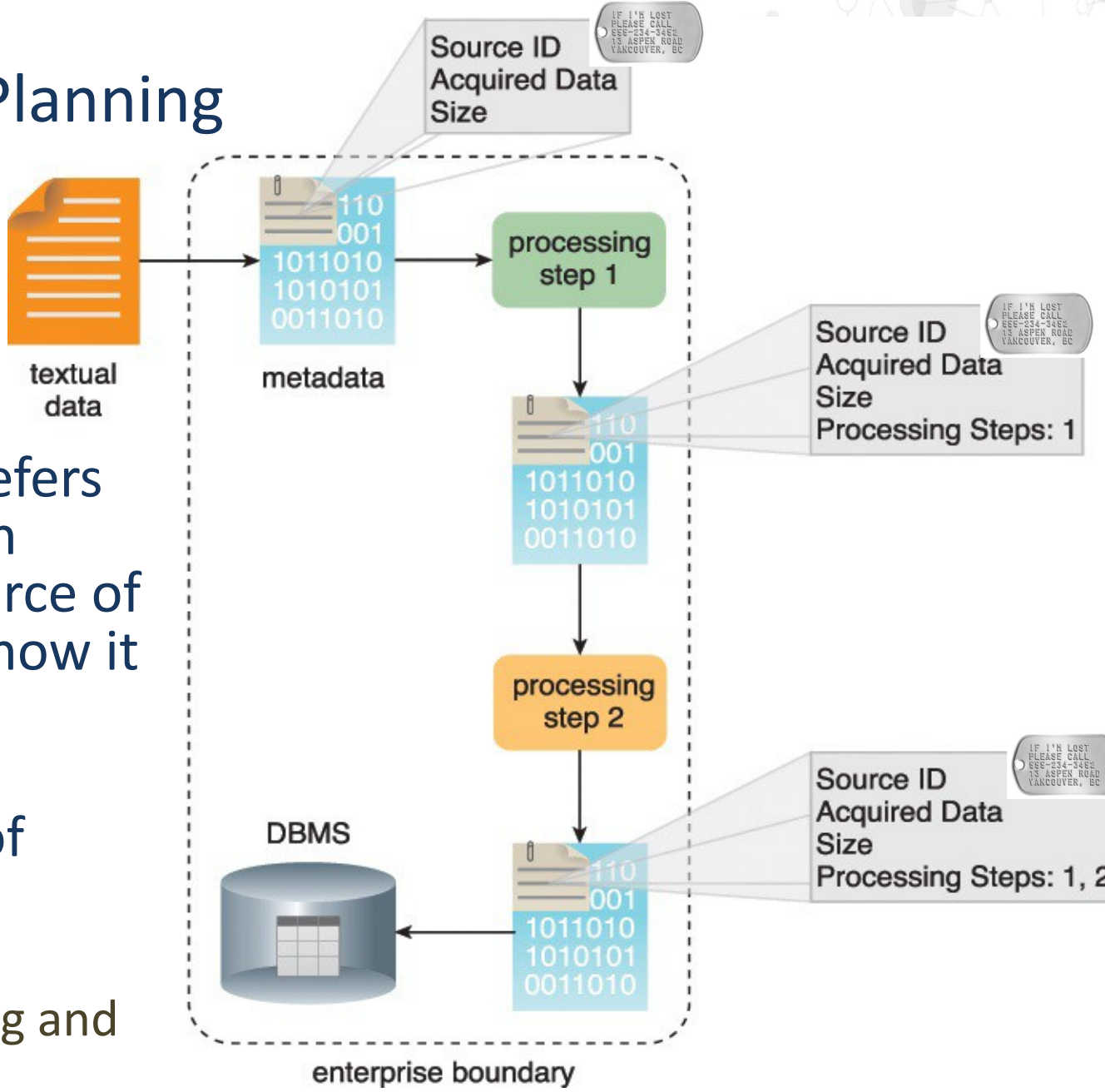
Data Security - Accumulo



Adoption and Planning

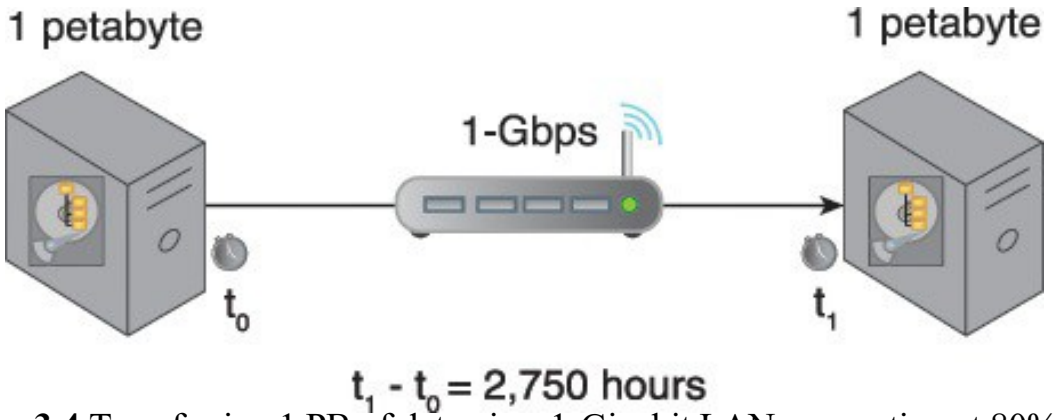
Provenance

- Provenance refers to information about the source of the data and how it has been processed.
- Two aspects of provenance include:
  - Data Tracking and Tagging



Big Data Planning Considerations

- 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.
- Distinct Performance Challenges
  - Due to the volumes and temperatures of data, some Big Data solutions are required to process data in real-time and performance is often a concern.



- **Latency** is the amount of time it takes to travel through the tube.
- **Bandwidth** is how wide the tube is.
- The amount of water flow will be your **throughput**

Figure 3.4 Transferring 1 PB of data via a 1-Gigabit LAN connection at 80% throughput will take approximately 2,750 hours

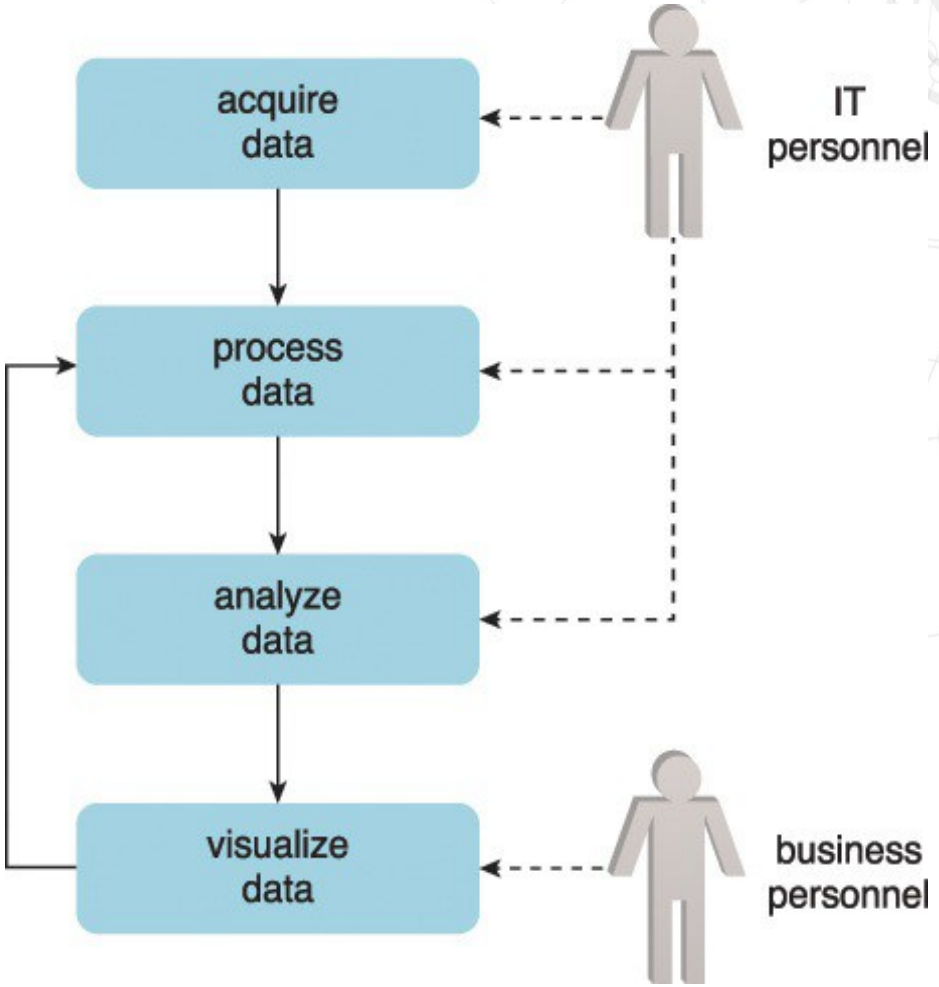
Big Data Planning Considerations

- 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**.
  - Chief Data Officer - SSORs and ADSs and Data Quality
  - Examples of what a Big Data governance framework can encompass include:
    - standardization of how data is tagged and the metadata within the tags
    - 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

Big Data Planning Considerations

Distinct Analytic Methodology

- A methodology will be required to control how data flows into and out of Big Data solutions.
  - DFaaS (Batch)
  - Pub/Sub (Warehouses)
- It will need to consider how feedback loops can be established to enable the processed data to undergo repeated refinement.
- "Human-in-the-loop"



Planning Considerations - Clouds

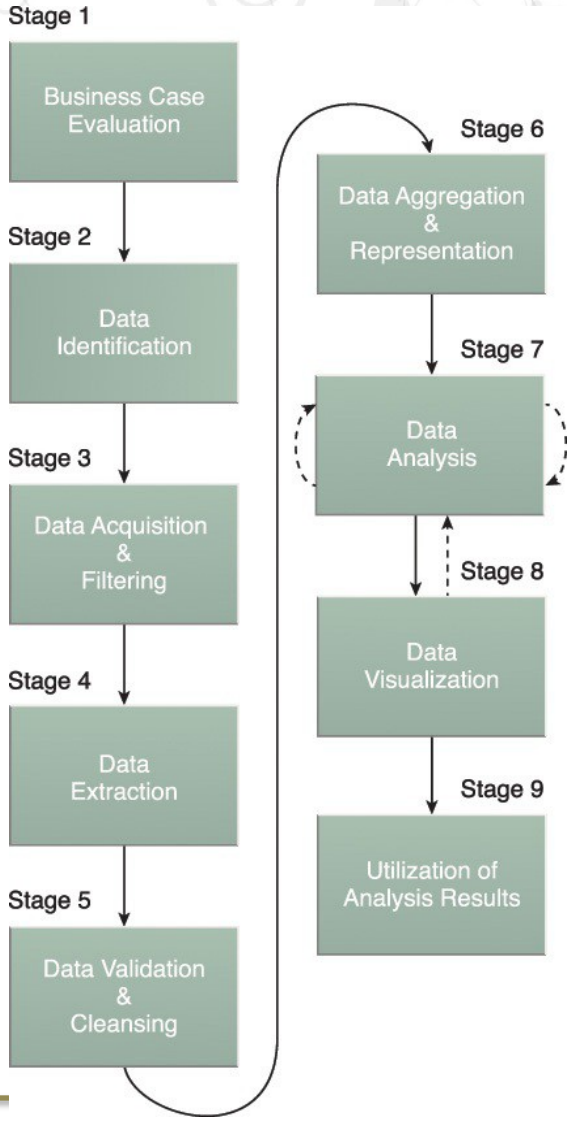
- Clouds provide remote environments that can host IT infrastructure for **large-scale storage and processing**, among other things.
- Common justifications for incorporating a cloud environment in support of a Big Data solution include:
  - inadequate in-house **hardware resources**
  - upfront capital **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 (e.g. private cloud)
  - 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

Big Data Analytics Lifecycle

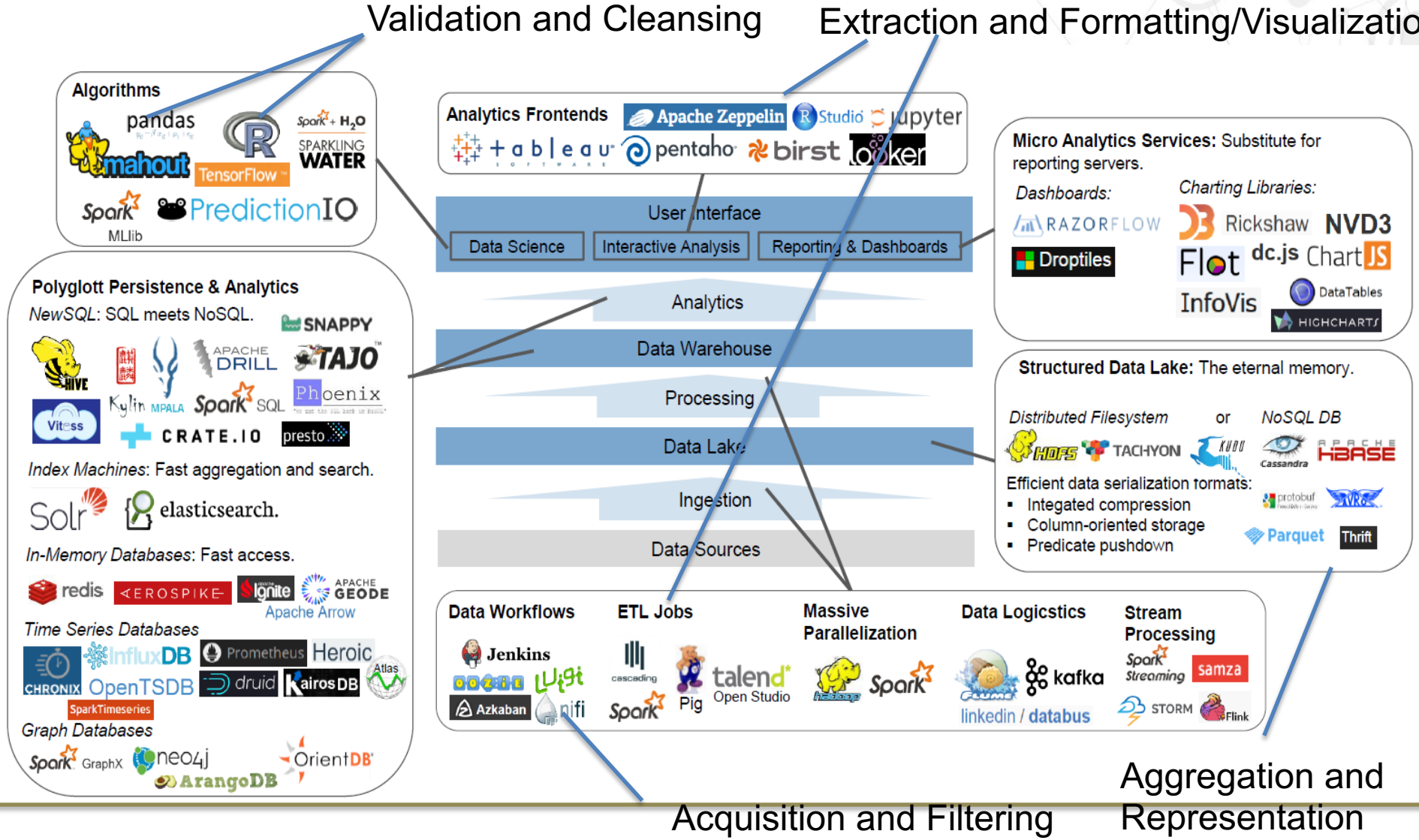
Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes.

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:

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

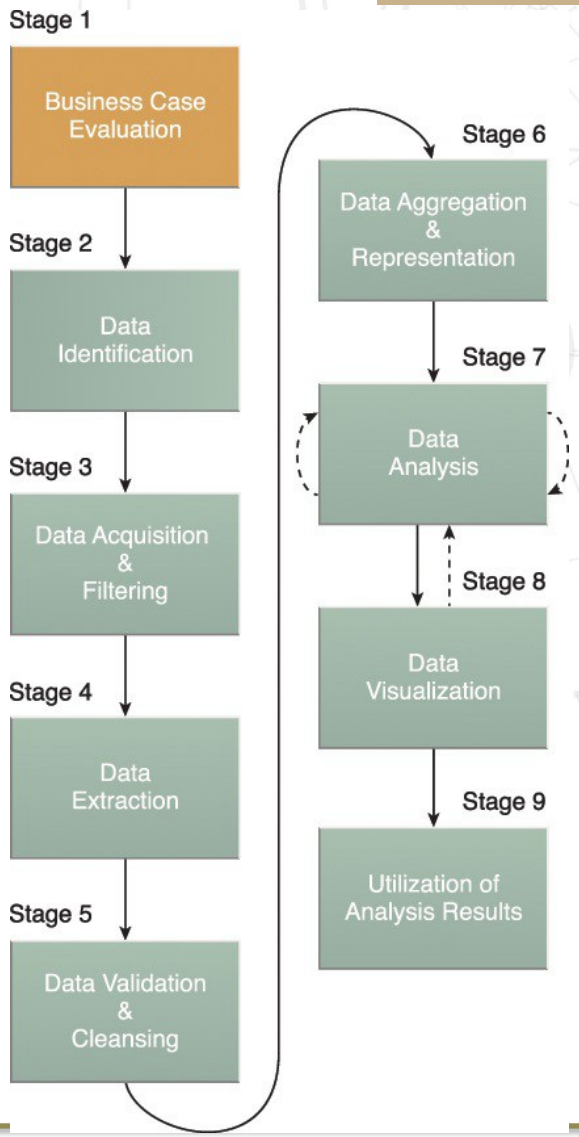


Data Life Cycle



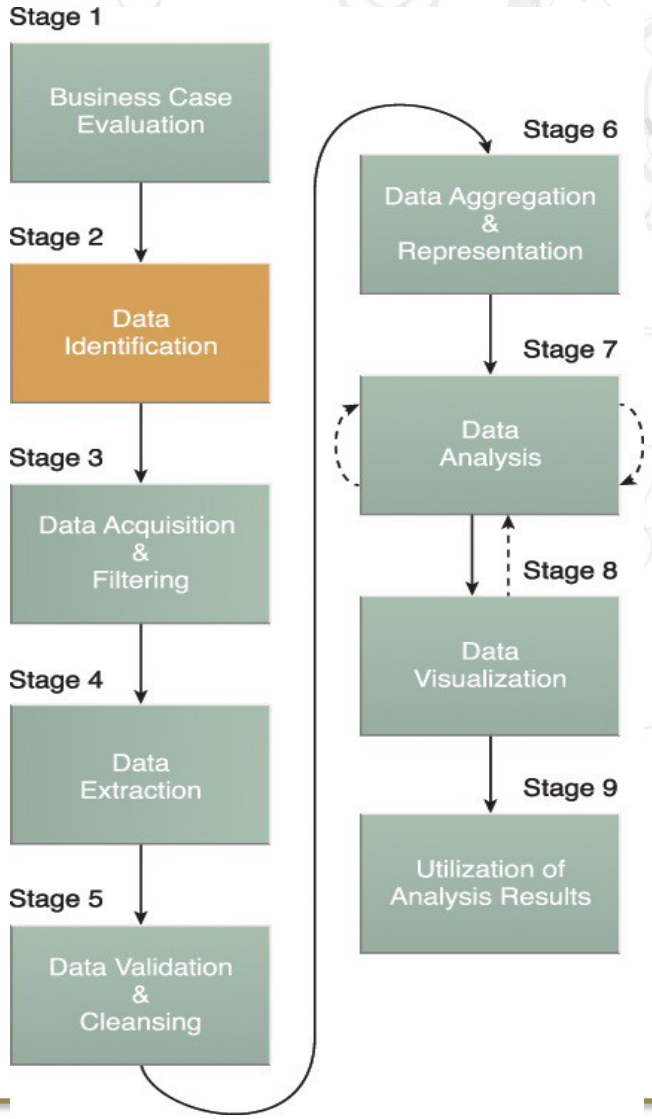
Business Case Evaluation

- Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the material processes and The Business **questions needed to conduct the analysis and how it is aligned with the goals of the organization.**
- Case Evaluation stage requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks (e.g. AVG).
- An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle (e.g. Solutions Process).



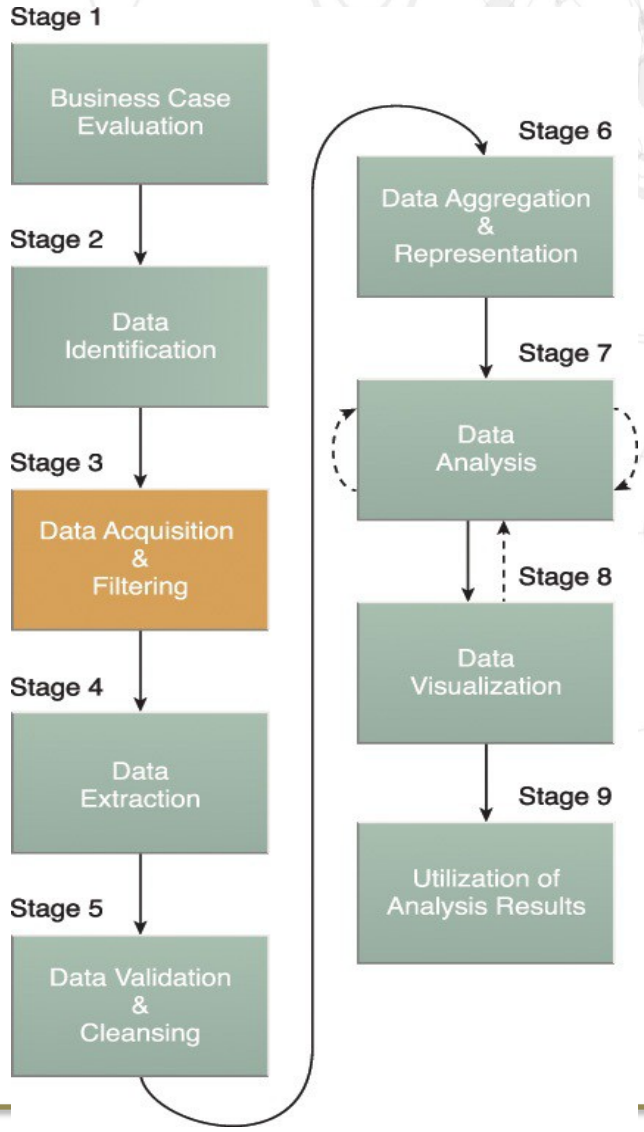
Data Identification

- The Data Identification stage is dedicated to **identifying the datasets** required for the analysis project and their sources.
- Identifying a **wider variety** of data sources may increase the probability of **finding hidden patterns and correlations.**
- Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be **internal and/or external** to the enterprise.



Data Acquisition & Filtering

- During the Data Acquisition and Filtering stage is the **gathering of data** from all sources identified during the previous stage (Data Identification).
- The acquired data is then subjected to automated **filtering** for the **removal** of **corrupt data** or data that has been deemed to have **no value** to the analysis objectives (e.g. NiFi).
- Depending on the type of data source, data may come as a collection of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter.



Data Acquisition & Filtering

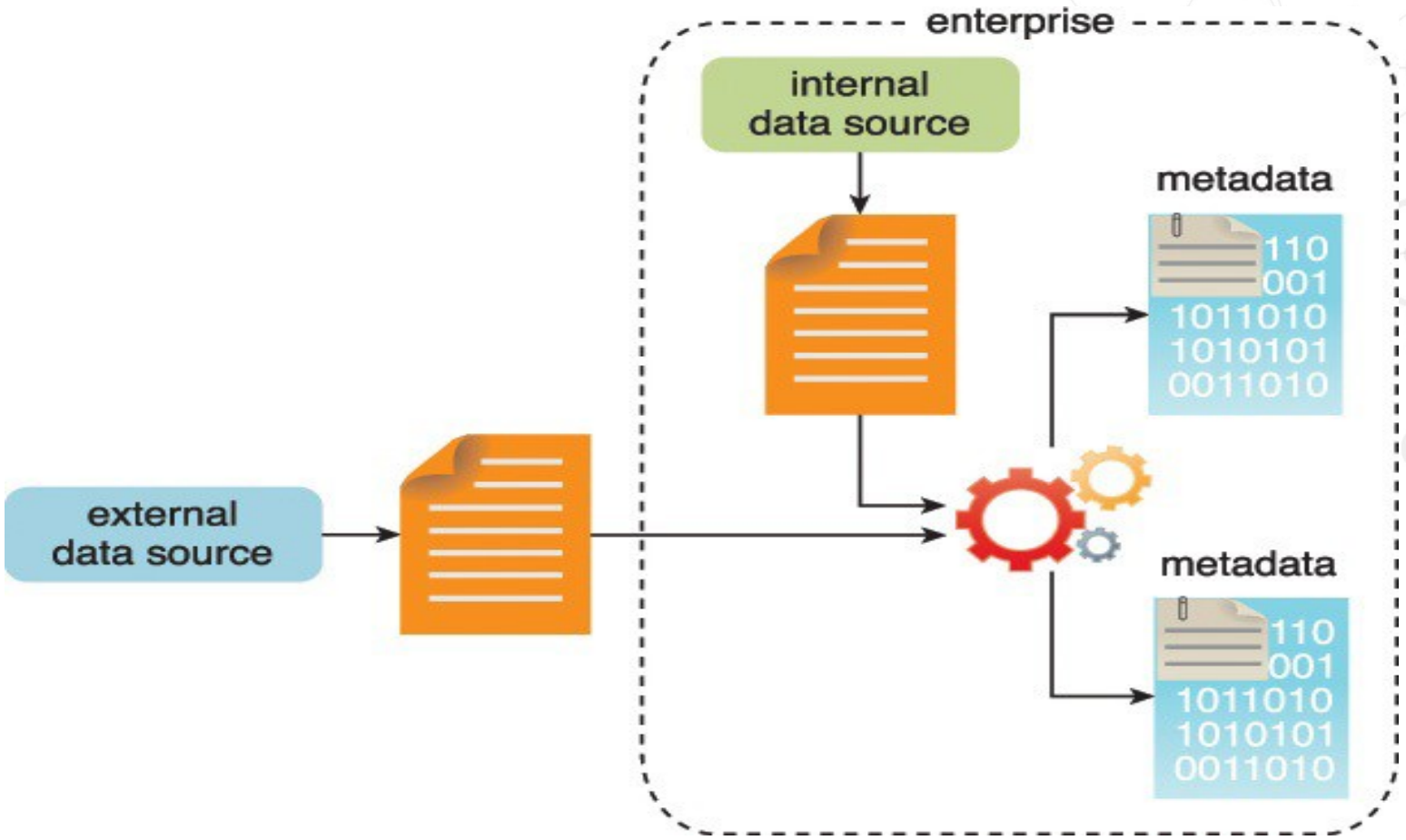
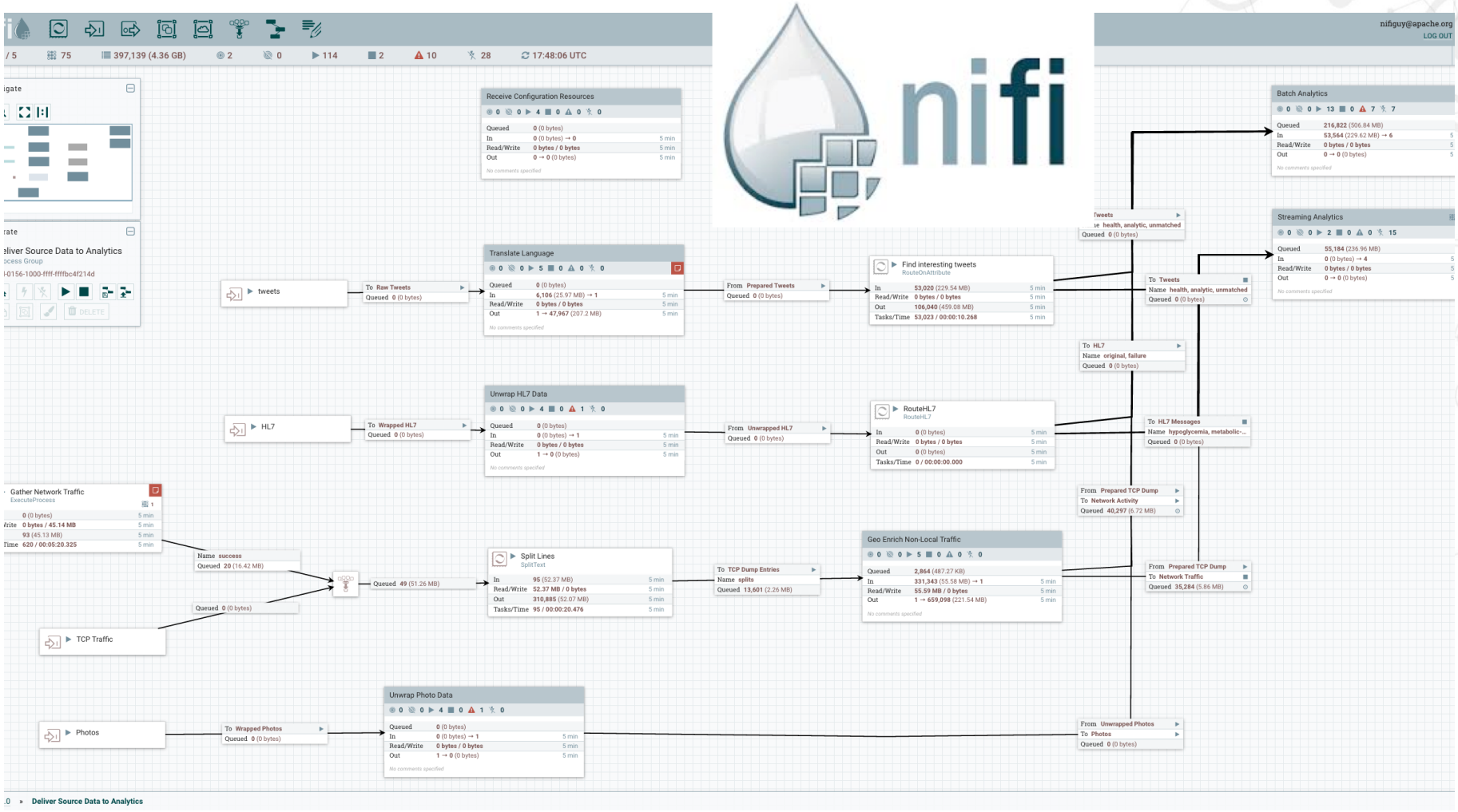


Figure 3.10 Metadata is added to data from internal and external sources.

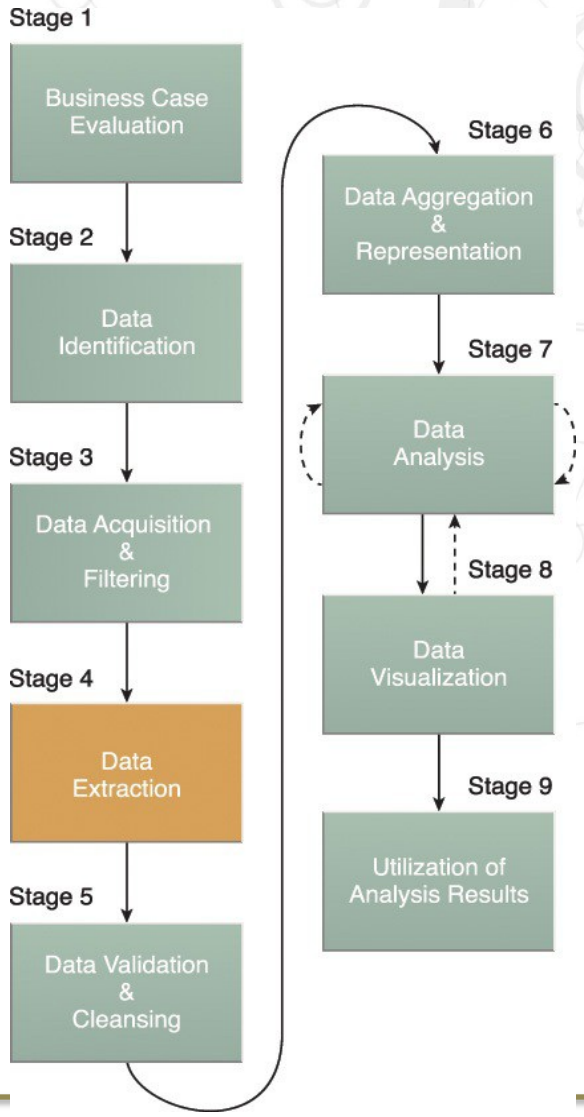
Data Acquisition and Filtering -NiFi



http://nifi.apache.org

Data Extraction & Formatting

- Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution.
- The Data Extraction lifecycle stage is dedicated to extracting disparate data and **transforming it into a format** that the underlying Big Data solution can use for the purpose of the data analysis (e.g. Data Format Standards - SOTF).
- The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution



Data Extraction

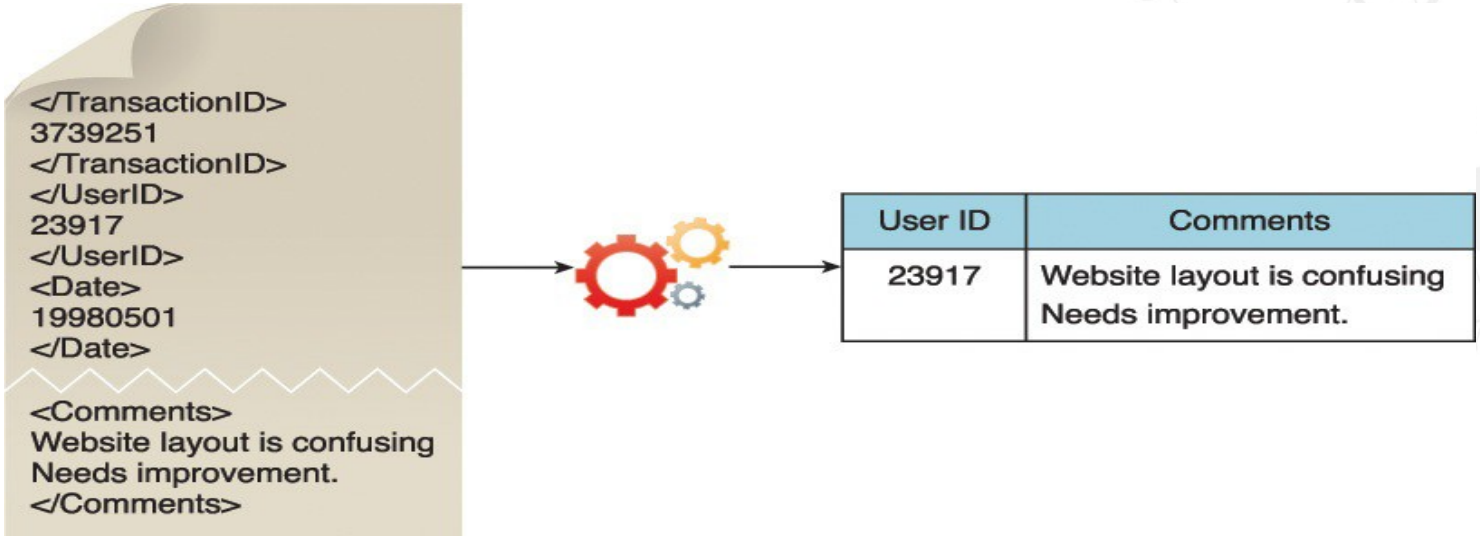


Figure 3.12 illustrates the extraction of comments and a user ID embedded within an XML document without the need for further transformation.

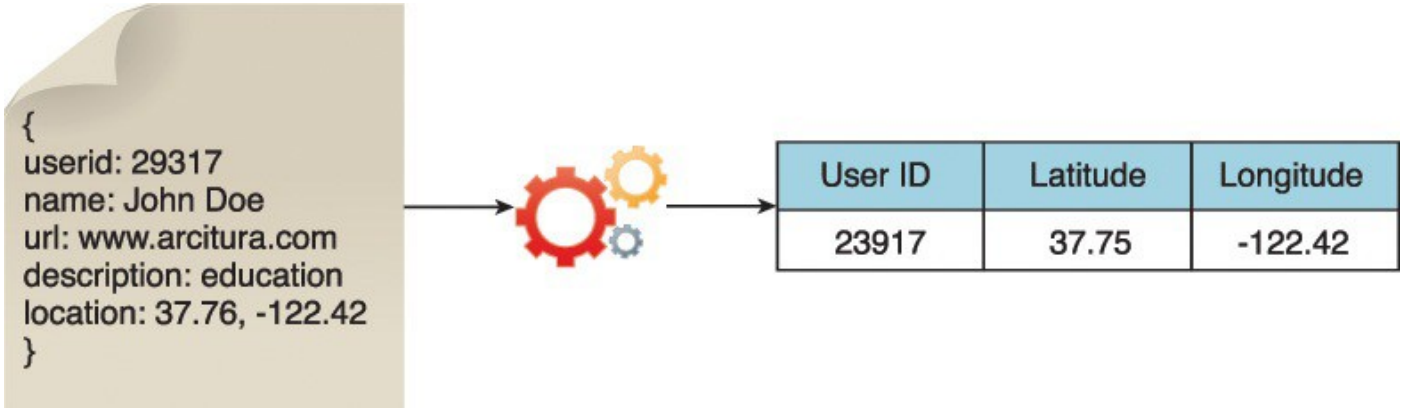
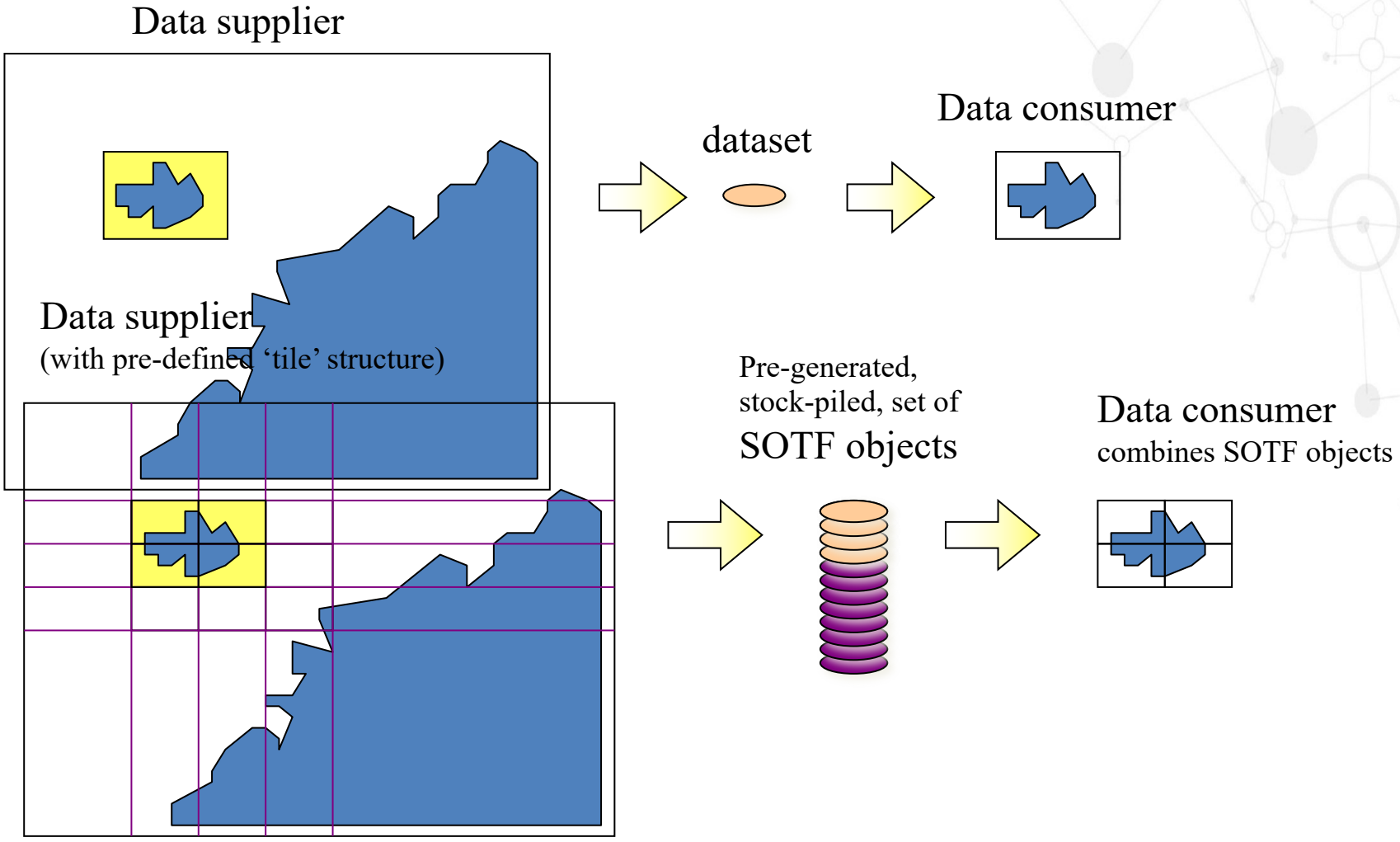


Figure 3.13 demonstrates the extraction of the latitude and longitude coordinates of a user from a single JSON field.

A Prototype Spatial Object Transfer Format (SOTF)

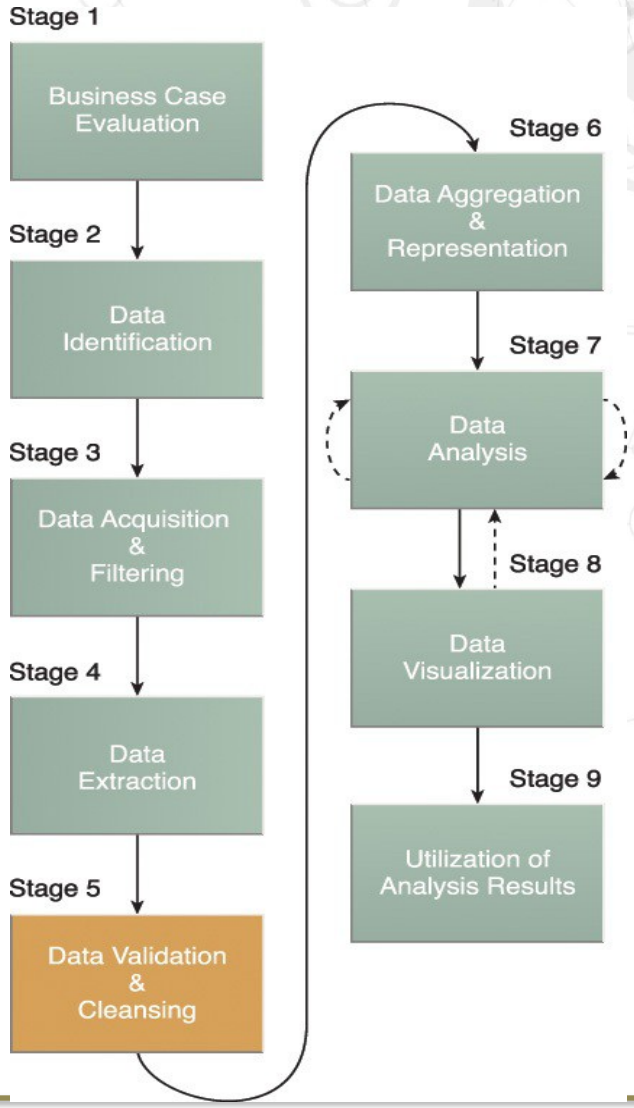
- SOTF supports multiple inheritance of feature types
- SOTF supports light-weight, binary feature relationships
- SOTF it was originally designed to handle complex, structured geospatial data;
  - *it does not support methods and behaviour.*
- SOTF has an object-oriented schema with:
  - features and feature types
  - properties and data types
- SOTF supports multiple geometric properties per feature
- SOTF supports both spatial and aspatial feature types
- Designed to work with both [object-]relational and object-oriented data stores
- An SOTF dataset always includes an *explicit* schema
- To support export of an SOTF dataset a data store
  - should provide feature identifiers that persist between exports
  - may provide ability to retain a previous state

A Prototype Spatial Object Transfer Format (SOTF)



Data Validation & Cleaning

- Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, **data input into Big Data analyses can be unstructured without any indication of validity**.
- This stage is dedicated to establishing often complex validation rules and removing any known invalid data.
- Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore **interconnected datasets** in order to assemble validation parameters and fill in missing valid data.



Data Validation & Cleaning

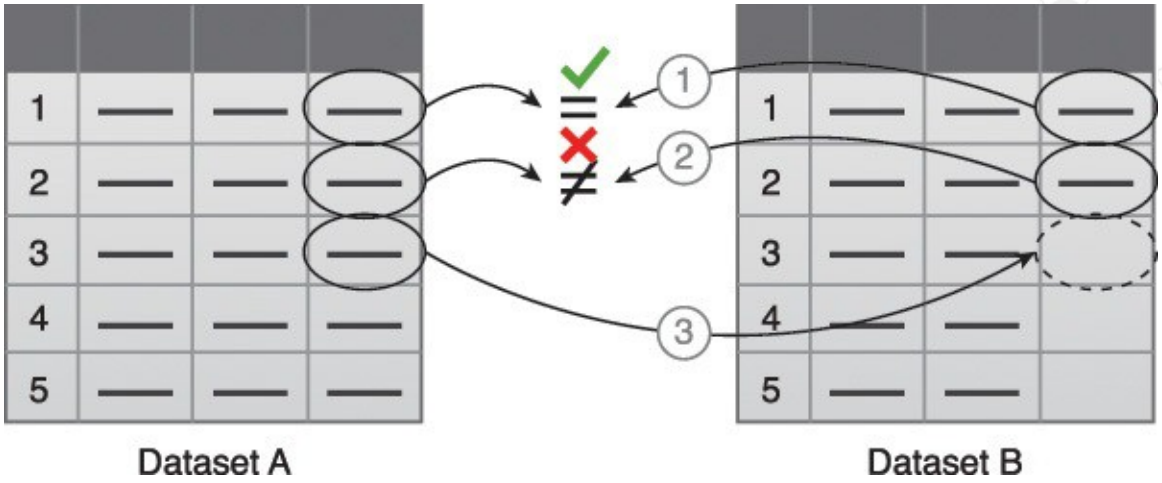


Figure 3.15 Data validation can be used to examine interconnected datasets in order to fill in missing valid data.

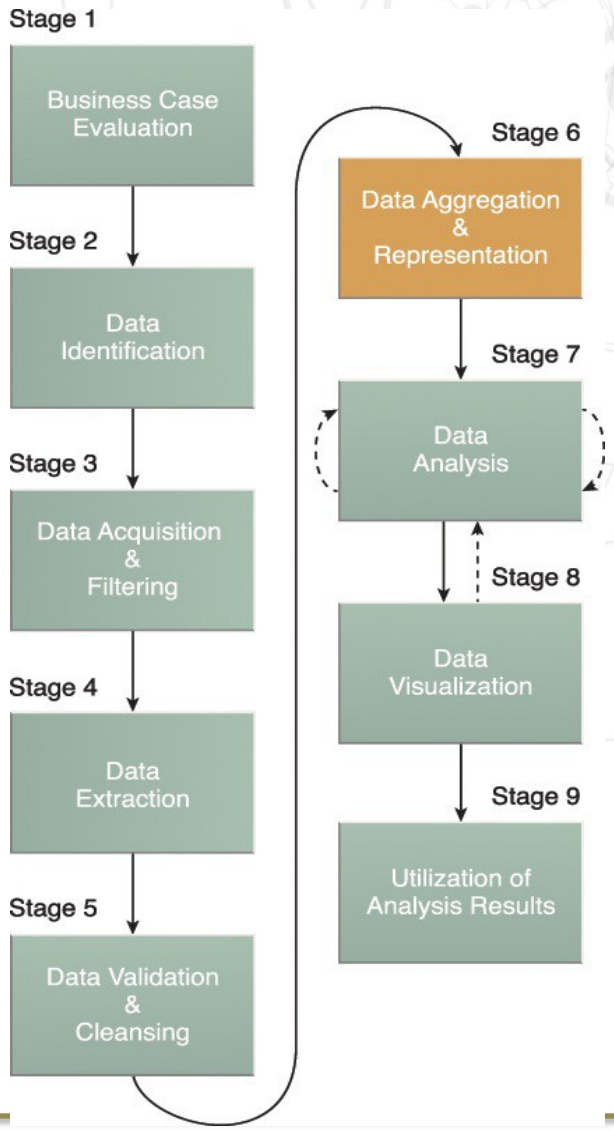
Figure 3.16 The presence of invalid data is resulting in spikes. Although the data appears abnormal, it may be indicative of a new pattern.

Data Aggregation & Representation

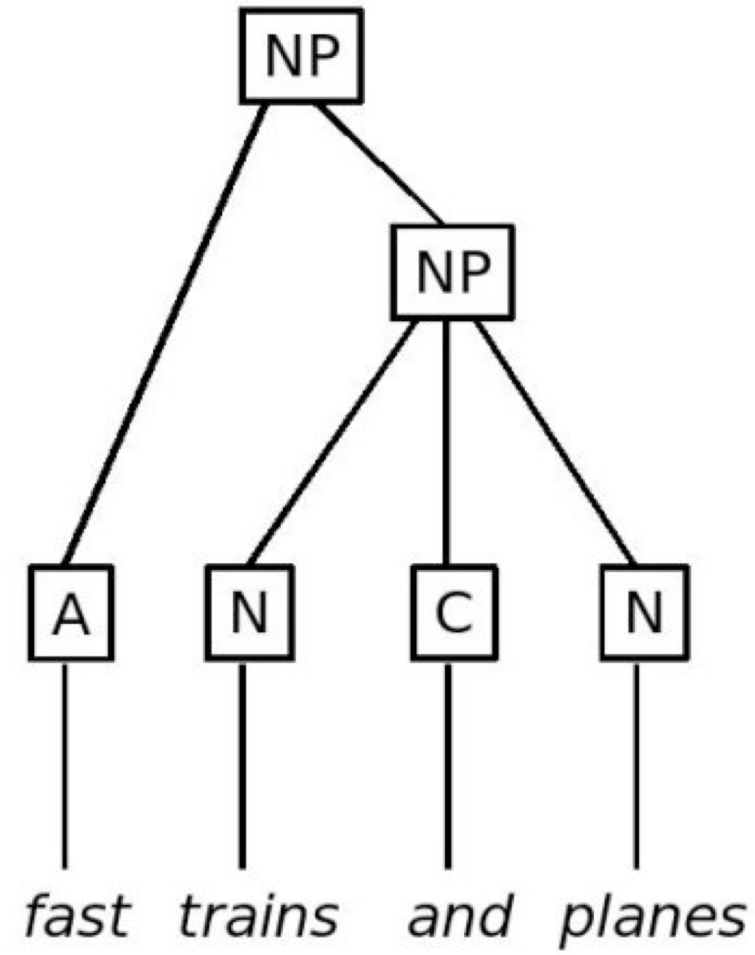
The Data Aggregation and Representation **stage is dedicated to** integrating multiple datasets together to arrive at **a unified view**.

Performing this stage can become complicated because of differences in:

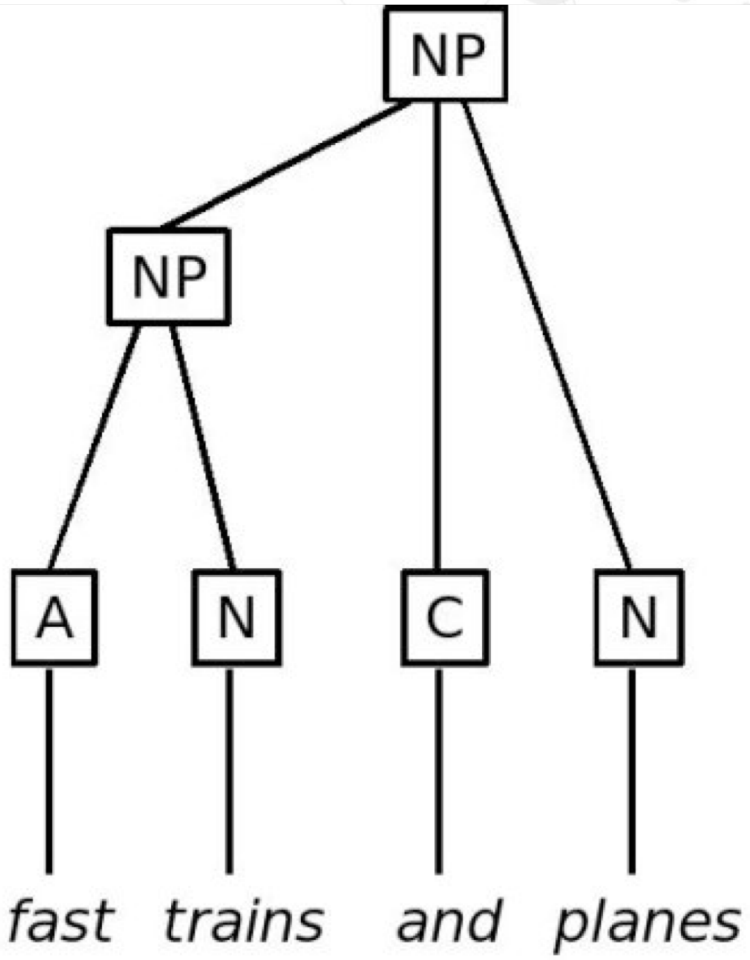
- ***Data Structure*** – Although the data format may be the same, the data model may be different.
- ***Semantics*** – A value that is labeled differently in two different datasets may mean the same thing, for example "surname" and "last name."



Data Aggregation & Representation



Interpretation A



Interpretation B

Data Aggregation & Representation

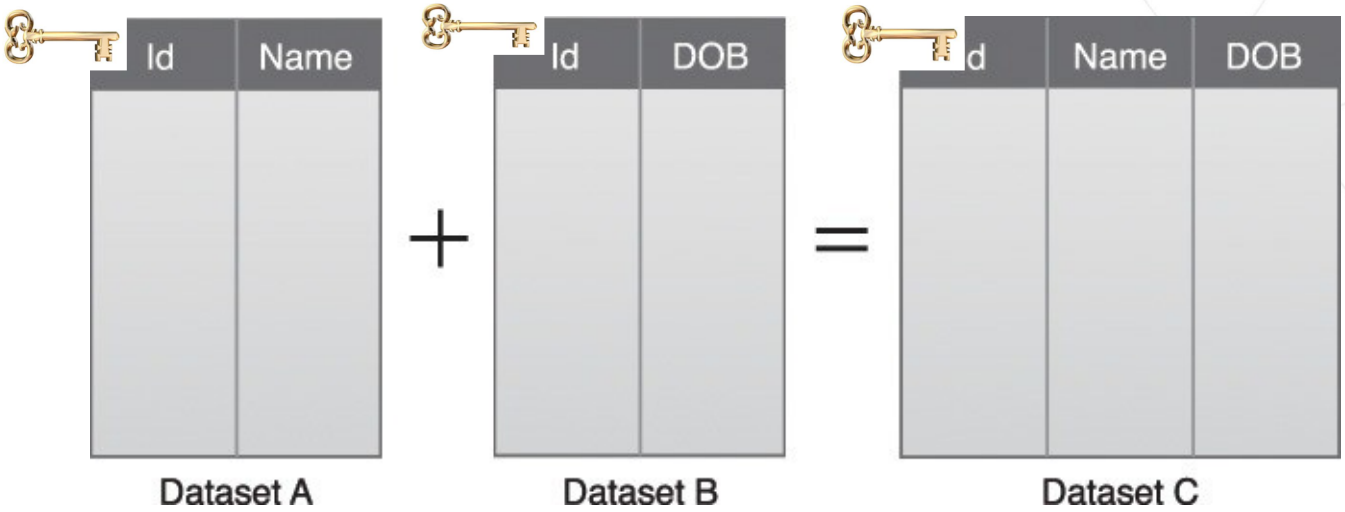


Figure 3.18 A simple example of data aggregation where two datasets are aggregated together using the Id field

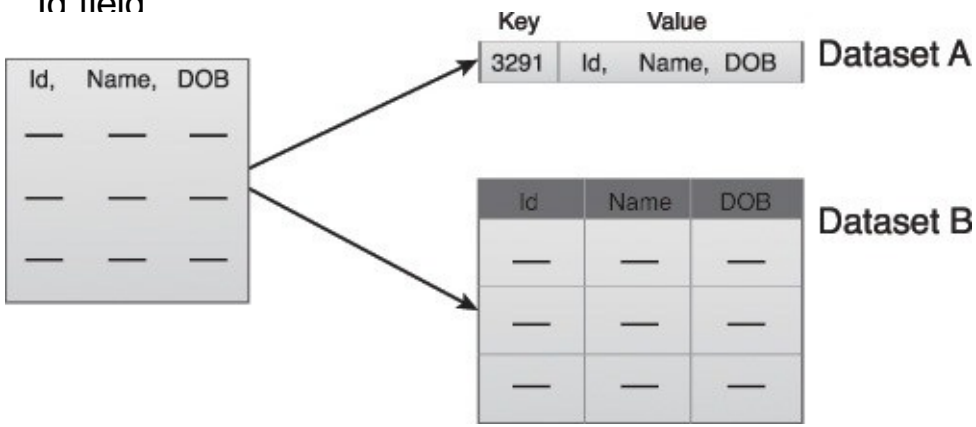
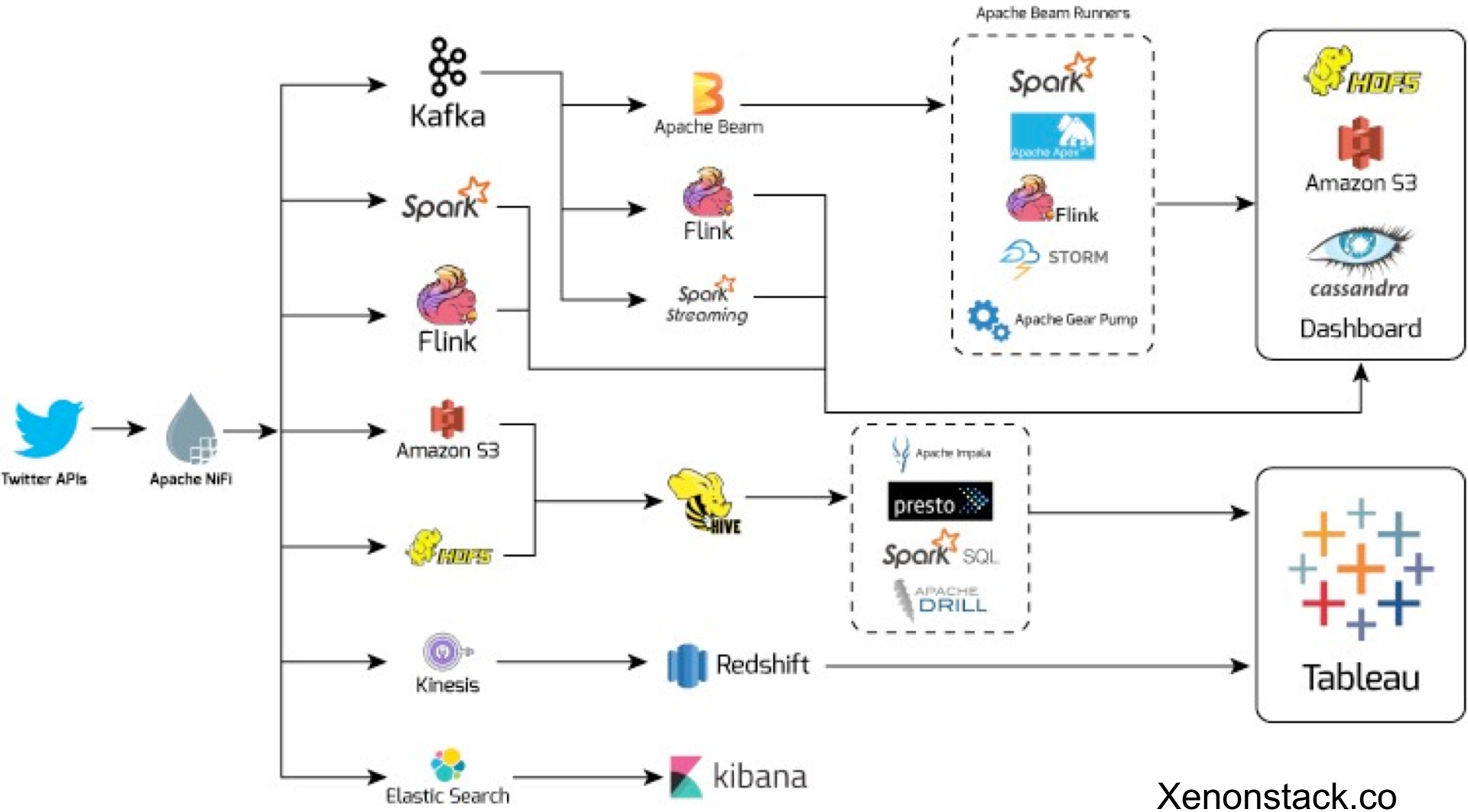


Figure 3.19 Dataset A and B can be combined to create a standardized data structure with a Big Data solution.

Building a Data Lake with NiFi



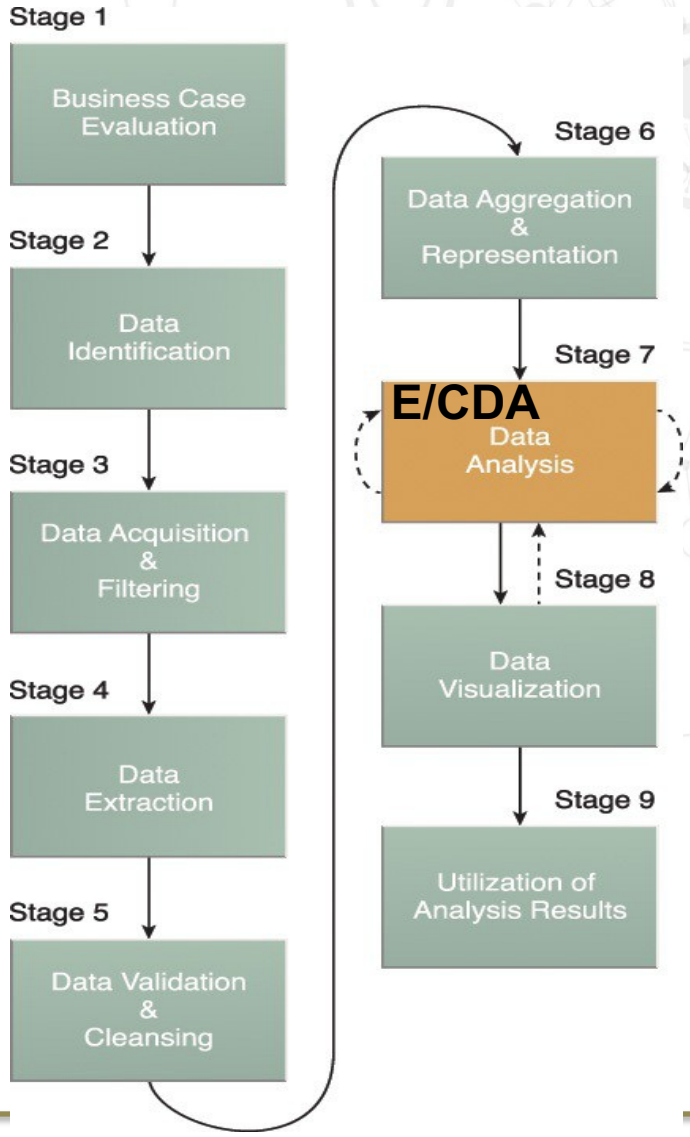
Xenonstack.com

Data Analysis

- The Data Analysis stage is dedicated to carrying out the actual **analysis task**, which typically involves one or more types of analytics.
- This stage can be **iterative** in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered.
  - The Exploratory Data Analysis (EDA) approach will be explained shortly, along with confirmatory analysis.

Type examples:

- querying a dataset to compute an **aggregation** for comparison
- **combining data mining and complex statistical analysis techniques** to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.



Data Analysis

- Data analysis can be classified as confirmatory analysis or exploratory analysis. The latter of which is linked to data mining
  - Data mining is the computing process of **discovering patterns in large data sets** involving methods at the intersection of machine learning, statistics, and database systems.
- **Confirmatory Data Analysis (CDA)** is a deductive approach where the cause of the phenomenon being investigated is proposed beforehand. The proposed cause or assumption is called a hypothesis.
- **Exploratory Data Analysis (EDA)** is an inductive approach that is closely associated with data mining. No hypothesis or predetermined assumptions are generated. Instead, the data is explored through analysis to develop an understanding of the cause of the phenomenon.

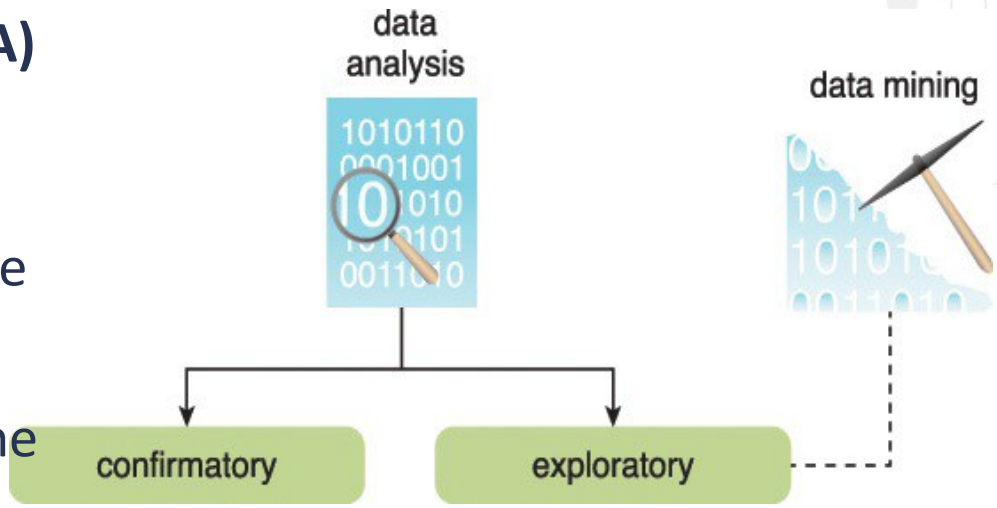
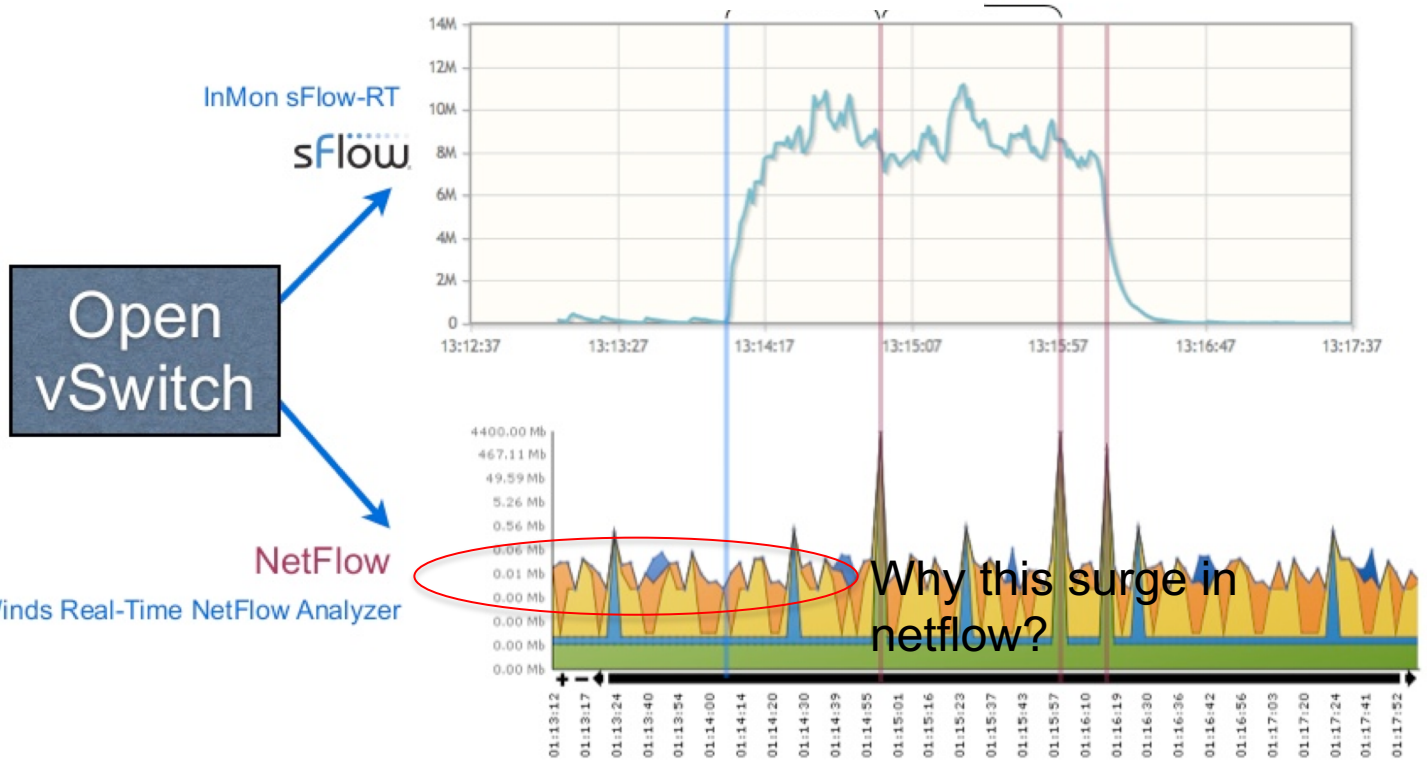


Figure 3.21 Data analysis can be confirmatory or exploratory analysis.

EDA and CDA

Just explore the data to see what is there.

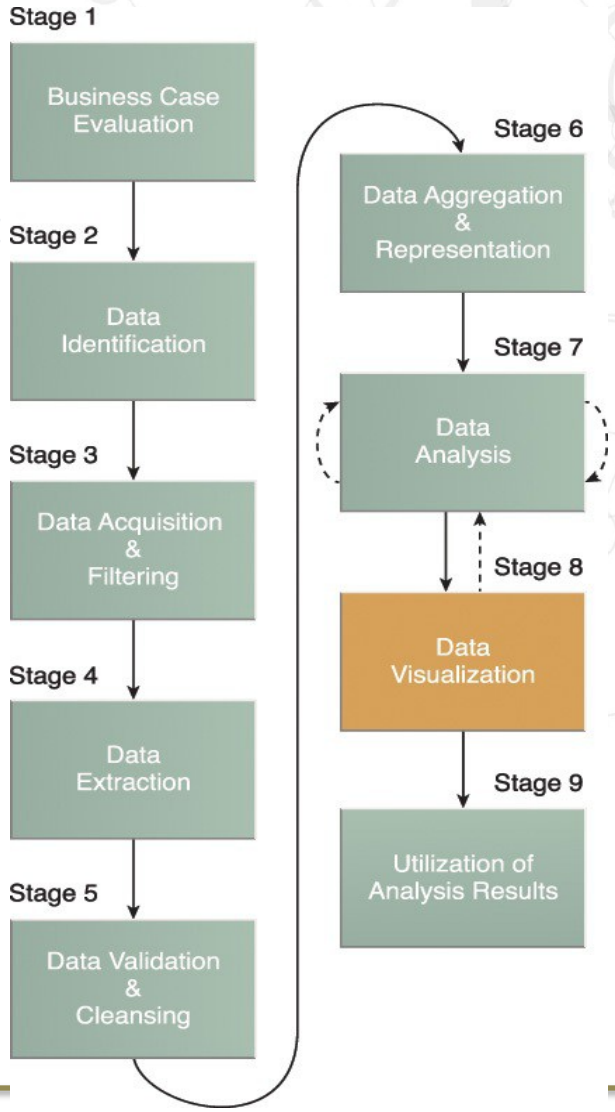


- sFlow does not use flow cache, so realtime charts more accurately reflect traffic trend
- NetFlow spikes caused by flow cache active-timeout for long running connections

Source: sflow.com

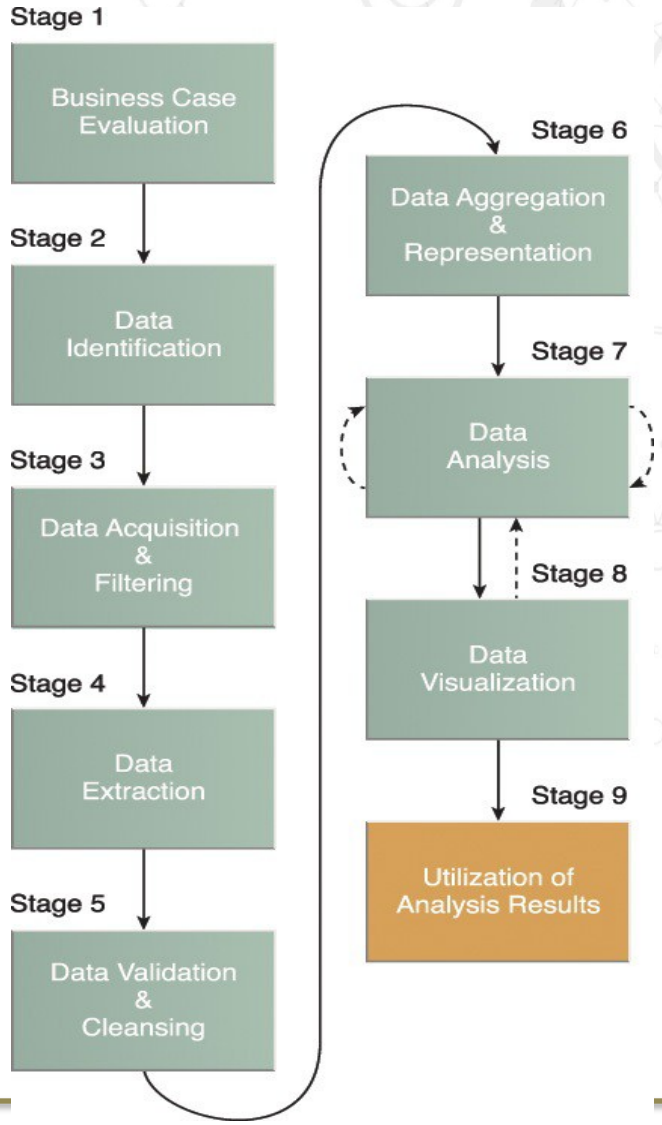
Data Visualization

- The Data Visualization stage is dedicated to **using data visualization techniques and tools to graphically communicate, represent and interpret** the analysis results for effective interpretation by business users.
- The Data Visualization stage is dedicated to using data visualization techniques and tools to **graphically communicate the analysis** results for effective interpretation by business users.
- The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated.



Utilization of Analysis Results

- Subsequent to analysis results being made available to business users to **support business decision-making,** such as via dashboards, there may be further opportunities to utilize the analysis results.
- The Utilization of Analysis Results stage is dedicated to **determining how and where processed analysis data can be further leveraged.**
- Depending on the nature of the analysis problems being addressed, it is possible for the analysis results **to produce "models**" that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed.



Select One of the Use Cases Below to answer your questions above.

1. **Log Analytics in Financial Services**

Banks use regular financial risk model calculations across all of their lines of business to determine their overall risk profile. Banks' IT teams run these calculations continuously to ensure that they comply with rules and that liquidity and cash balances are adequate. In order to sufficiently analyze and monitor application performance in real-time to ensure a near 100% service level requirement for these critical applications, banks can depend on open source log analytics solutions.

2. **E-Commerce**

A system-wide logging infrastructure combined with log analytics is a powerful way to improve operational performance in IT and business. By analyzing event logs and system metrics that track the performance of IT systems and user logs that capture the behavior of users who interact with e-commerce sites, retailers can gather data and discover insights leading to more agile operations, improved competitive advantage, and increased site revenue.

3. **Market Research**

Market research is a very important component of today's business strategy. Market research techniques typically encompass the analysis of data to gain insight or to support decision making. However, many market research firms have not improved their market research collection and analysis processes for years. In addition, both the amount of data and the number of data sources have increased exponentially in recent years. Thus, leveraging big data analytics solutions and scalable data processing techniques allows these firms to gather valuable takeaways and insights that could immensely benefit their subscribers.

4. **Precision Agriculture**

Agriculture has molded human history for centuries. Agricultural technology advancements have made collecting and storing data easy; but what about processing and analyzing this data to support productivity? How do we process and obtain information from all this data, which keeps growing in real-time? Combining search engines, big data, and analytics, we now have a method to tie agricultural data together, making agriculture an exact science, saving costs, and increasing productivity.

5. **Precision Medicine**

The scale of genomics information being collected is enormous. Today's healthcare organizations and research institutes struggle with processing and extracting value out of the increasing deluge of genomics data triggered by the decreasing costs of genomics data sequencing. These organizations have increasingly been relying on big data analytics applications to ingest genomics data, physician notes, and medical research information in order to improve patient health, alleviate future health risks, and discover cures for diseases.