Before Lecture…

- This Go over Proposal Handout
- Portfolio – Resources: Dataset Listing and Tutorials
  - Next Week - 9/21 we will:
    - Portfolio Template (will try to set up in class)
    - Assignment 1 Instructions
- Student Research Proposal
  - Proposal Due: 9/28
- Next Week - Big Data Analysis, Technology Concepts, and Statistical Techniques

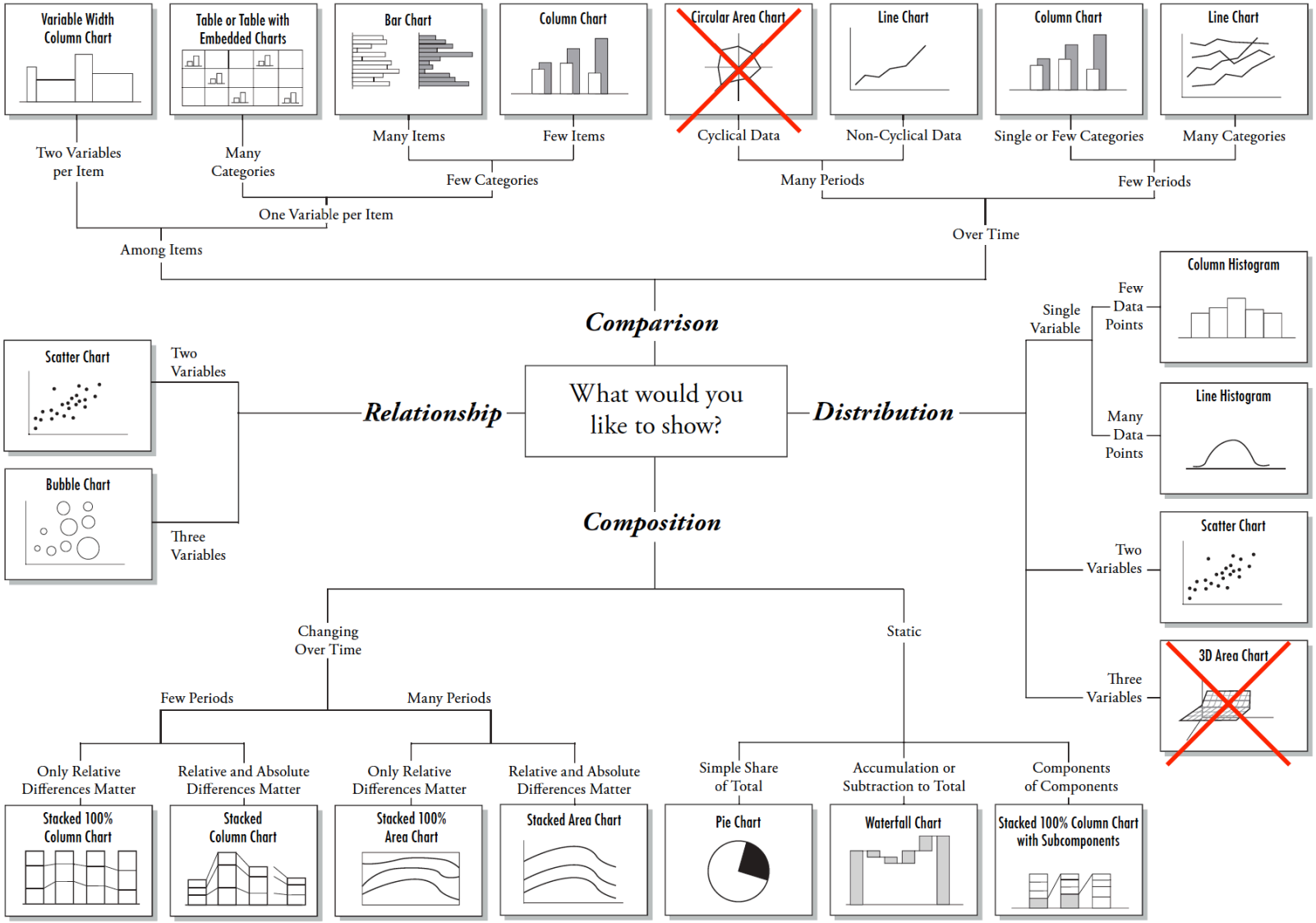
Proposal Instructions

- Take a look at the handout
- List the following in the Business Case Evaluation under the Methodology:
  - **Goals –** Questions
  - **Processes –** Analysis and Visualizations
    - Analysis - Python Statistical Libraries
    - Visualizations - Python Visualization Libraries
  - **Materials –** Data
  - **Technology –** Processing/Storage/etc.
    - E.g. Python Data Processing Libraries and Data Structures
- Example:
  - Goal: Are patients cancerous or benign based upon their genetic makeup?
  - Processes:
    - **Machine learning** - mlpy, scikit-learn – k-nn/k-means classification and clustering
    - **Visualization:** matplotlib, plotly – Relationships Bubble Chart
  - Materials: Global Cancer Map (GCM) data which contains…..
    - Source: http://globalcancermap.com
  - Technology
    - Storage: Python/HDFS - pandas
    - Processing: Python/Spark/MapReduce – nltk/gensim

Commonly Used Data Sciences Tools

Library	Usage
numpy, scipy	Scientific & technical computing
pandas	Data manipulation & aggregation
mlpy, scikit-learn	Machine learning
theano, tensorflow, keras	Deep learning
statsmodels, scrapy	Statistical analysis
nltk, gensim	Text processing
networkx	Network analysis & visualization
bokeh, matplotlib, seaborn, plotly	Visualization
beautifulsoup, scrapy	Web scraping

Visualization Types



Lecture: 3

Big Data Adoption and Planning Considerations Enterprise Technologies and Big Data Intelligence

Benjamin Harvey
Department of EMSE and CS
Data Analytics Program
The George Washington University
E-mail: bsharve@email.gw.edu
Web: https://bsharvey.github.io

Agenda – Planning and Considerations

- Organization Prerequisites: Big Data Frameworks
- Data Procurement
- Privacy
- Security
- Provenance
- Limited Real-time Support
- Performance Challenges
- Distinct Governance Requirements
- Distinct Methodology
- Clouds Big Data Analytics
- Data Analytics Lifecycle (analysis vs. analytic)

Agenda – Enterprise Technologies and Big data Business Intelligence

- Batch Processing
  - Online Analytical Processing (OLAP)
- Stream Processing
  - Online Transaction processing (OLTP)
- Extract Transform Load (ETL) Data
- Warehouses
- Data Marts
- Traditional Business Intelligence (BI)
- Big Data Business Intelligence (BI)
  - Understand the business side is key to success and this is what bringing in an Industry/Gov't prof. allows

Agenda 9/21 - Big Data Analysis, Technology Concepts, and Techniques

- Analysis Concepts: Models, EDA, CDA, Data Product, Statistics, Machine Learning, Data Munging,
- A/B Testing,
- Correlation and Regression
- Heat Maps
- Time Series Analysis
- Network Analysis
- Spatial Data Analysis
- Classification and Clustering
- Outlier Detection Filtering (including collaborative filtering & content-based filtering)
- Natural Language Processing
- Sentiment Analysis
- Text Analytics

Why Do I Care?

- 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
  - Data Science and Analytics
- Business, Applications, Information, Math & Technology (e.g. Research DS vs. Enterprise DS)

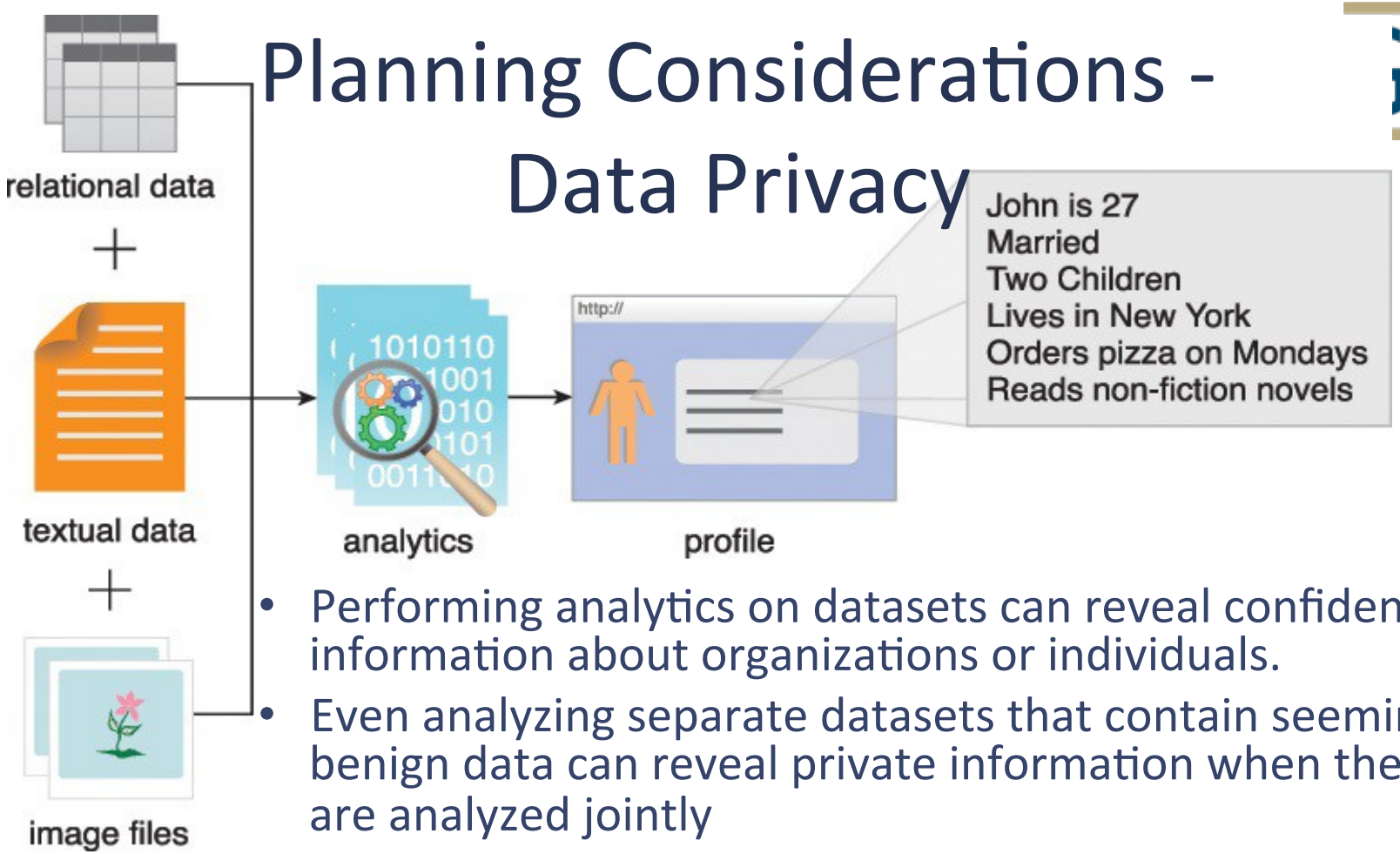
Big Data Planning Considerations

- 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.
    - https://hadoopecosystemtable.github.io/
  - External data sources include government data sources and commercial data markets.
  - Resources section of Portfolio
    - http://bsharvey.github.io

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.

Figure 3.1 Information gathered from running analytics on image files, relational data and textual data is used to create John's profile.

Joint Analytics and Privacy

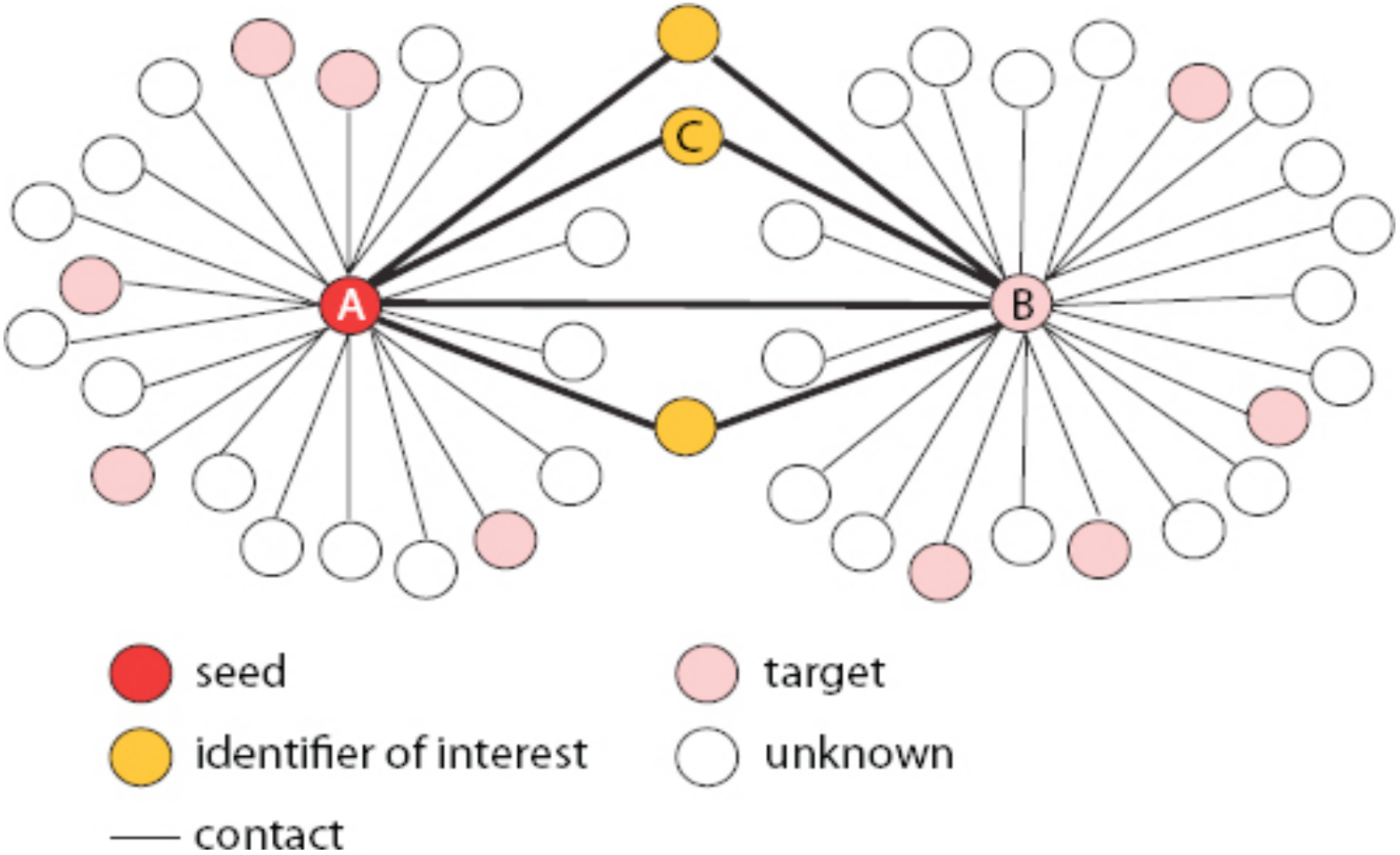


FIGURE 3.1 A network of contacts among identifiers
https://www.nap.edu/read/19414/chapter/5#43

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

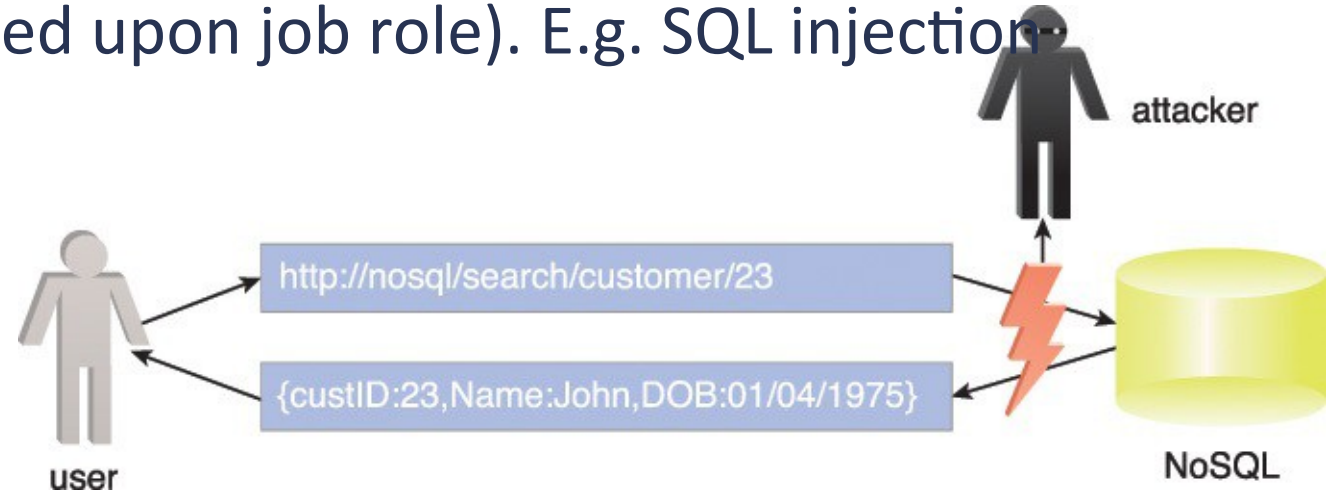
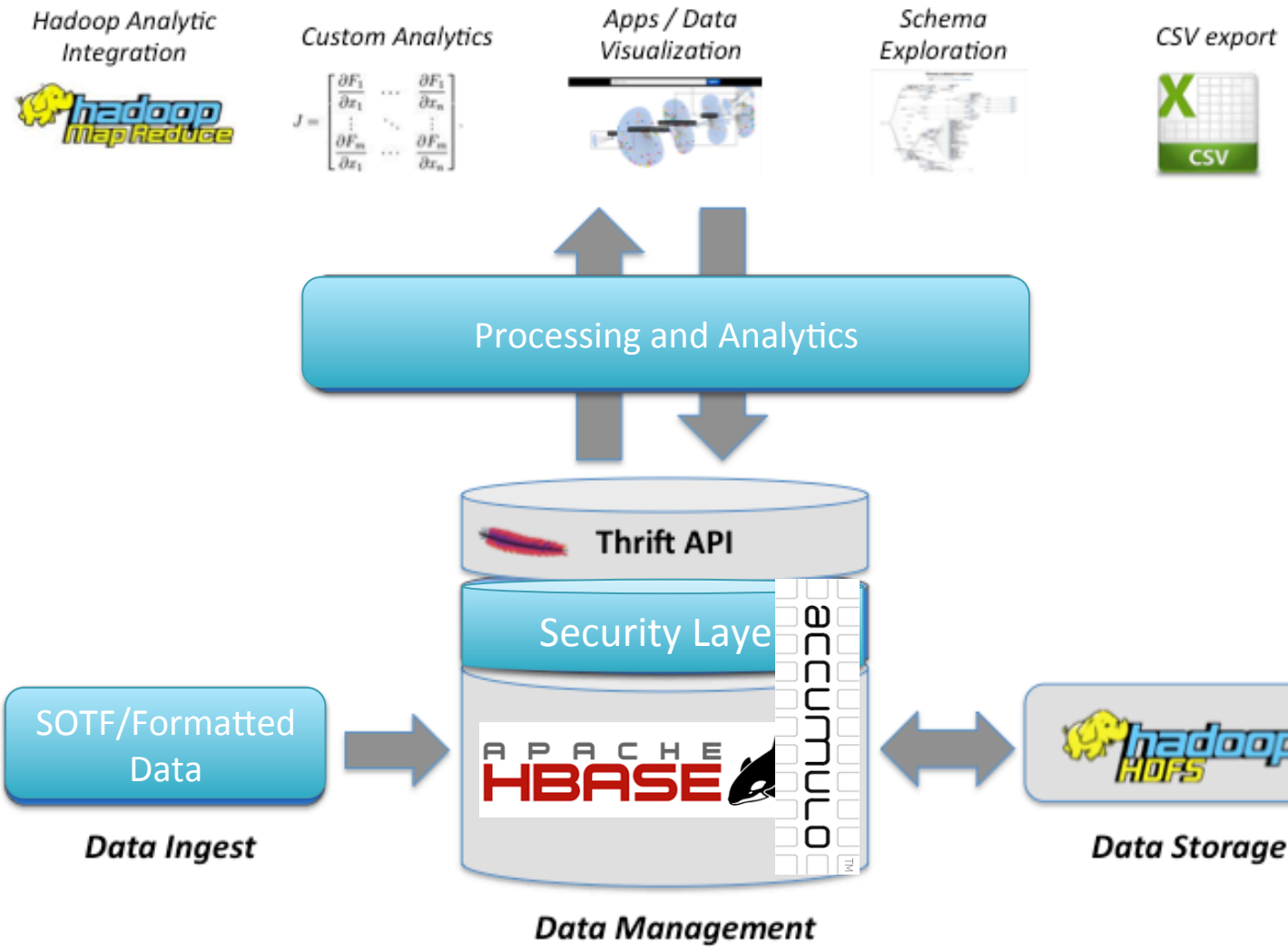


Figure 3.2 NoSQL databases can be susceptible to network-based attacks.

Data Security - Accumulo



Adoption and Planning Considerations

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

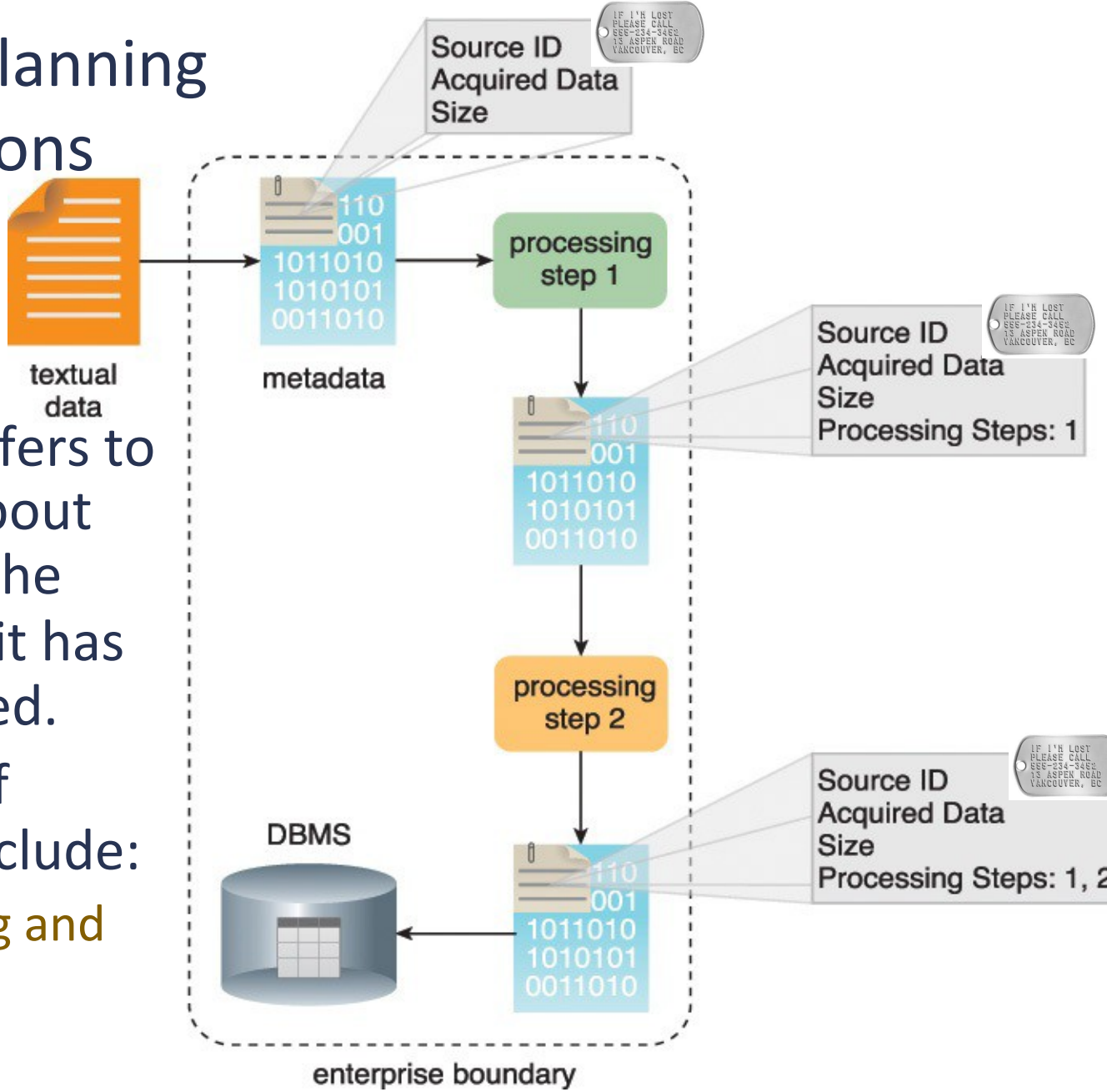
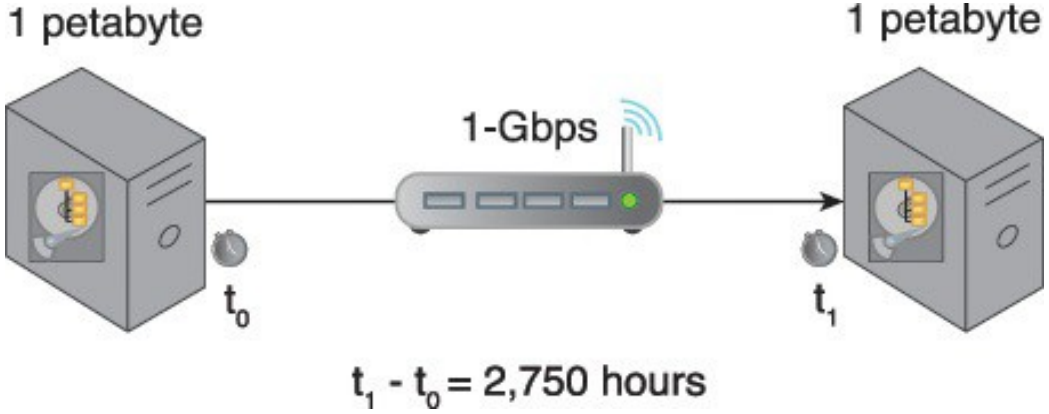


Figure 3.3 Data may also need to be annotated with source dataset attributes and processing step details as it passes through the data transformation steps.

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 of data that some Big Data solutions are required to process in real-time, 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

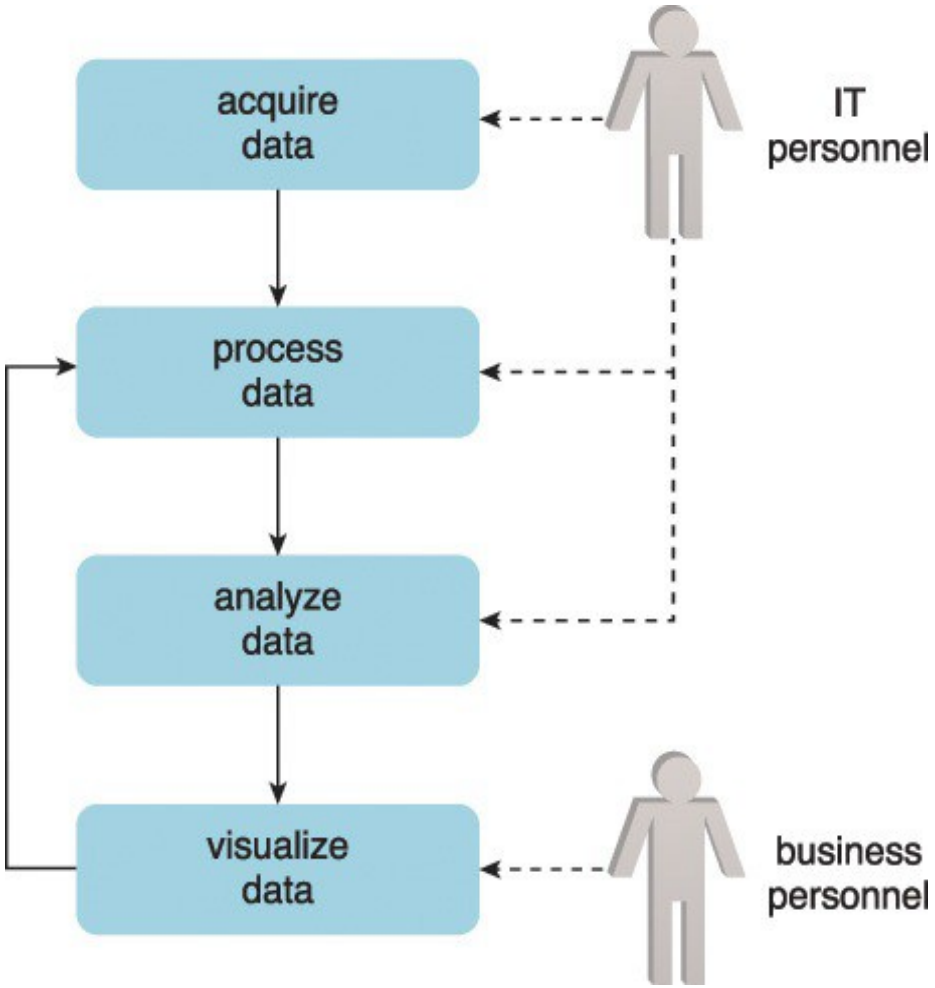
Big Data Planning Considerations

- 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**.
  - 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.
- 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

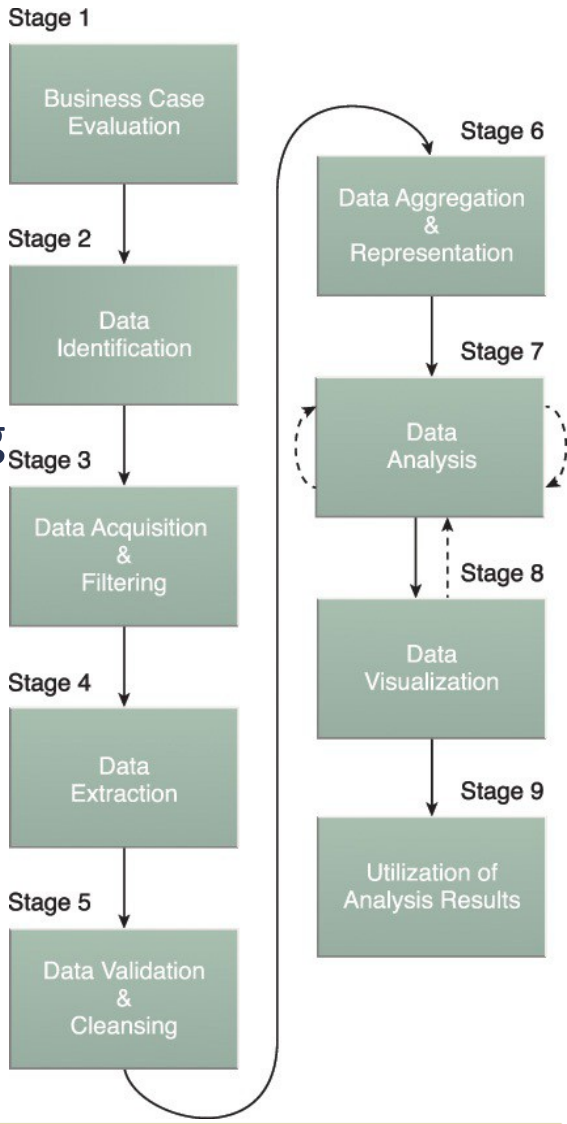
- Clouds provide remote environments that can host IT infrastructure for **large-scale storage and processing**, among other things.
- Common justifications for incorporating a cloud environment in support of a Big Data solution include:
  - inadequate in-house **hardware resources**
  - upfront capital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

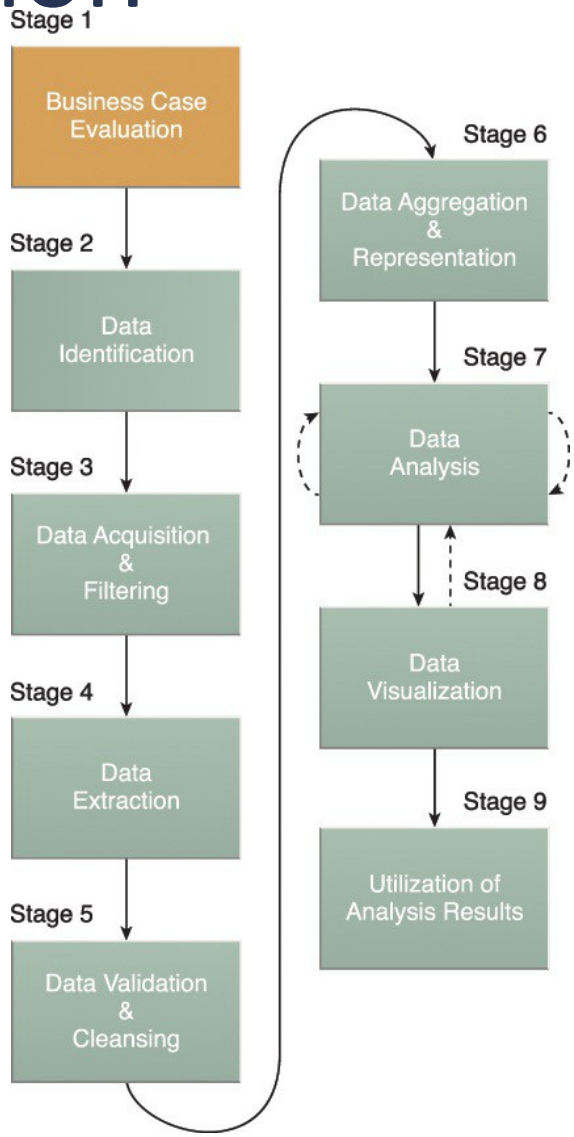
To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data:

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



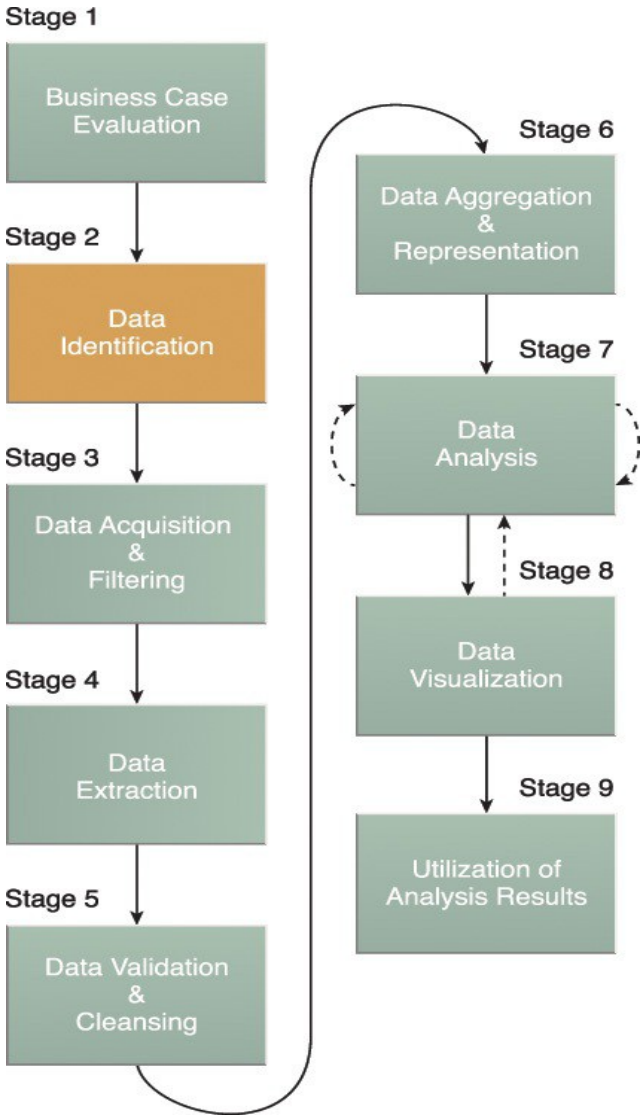
Business Case Evaluation

- Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the material processes and questions needed to conduct the analysis and how it is aligned with the goals of the organization.
- The Business 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).



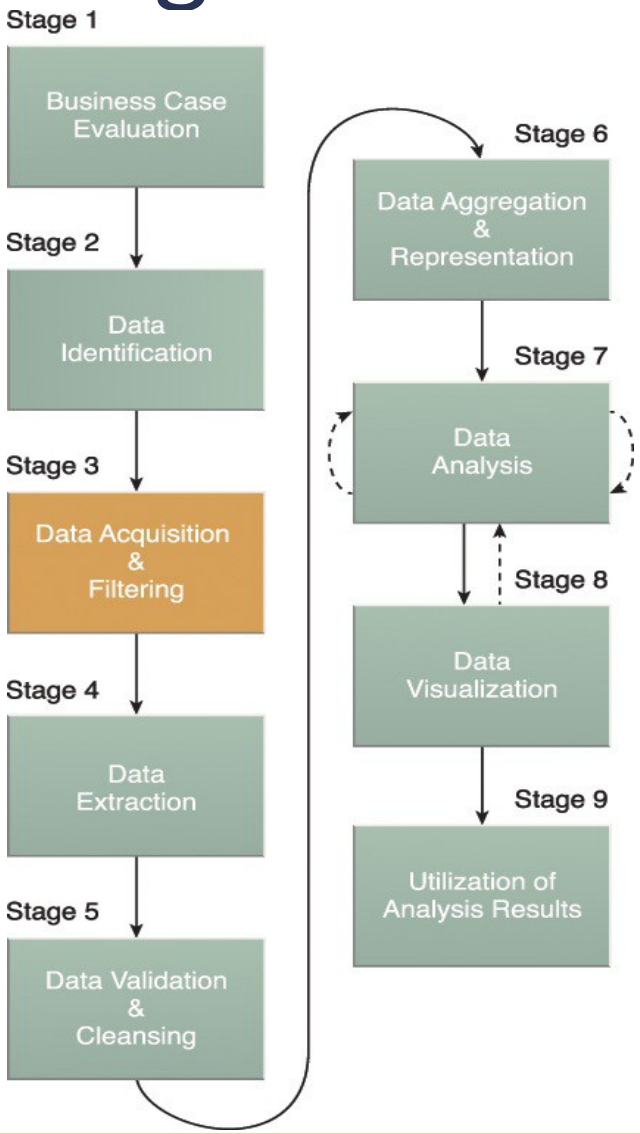
Data Identification

- 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

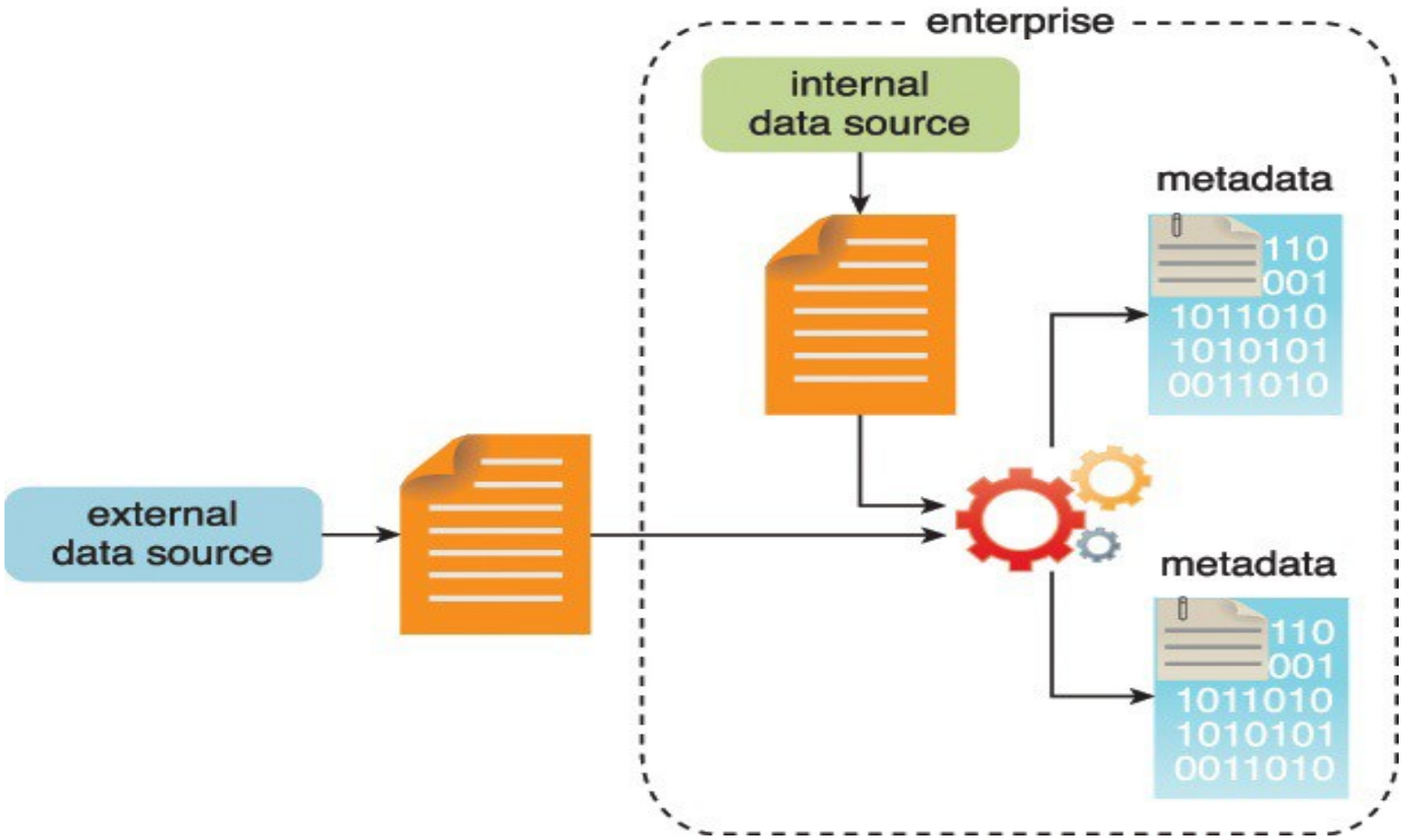
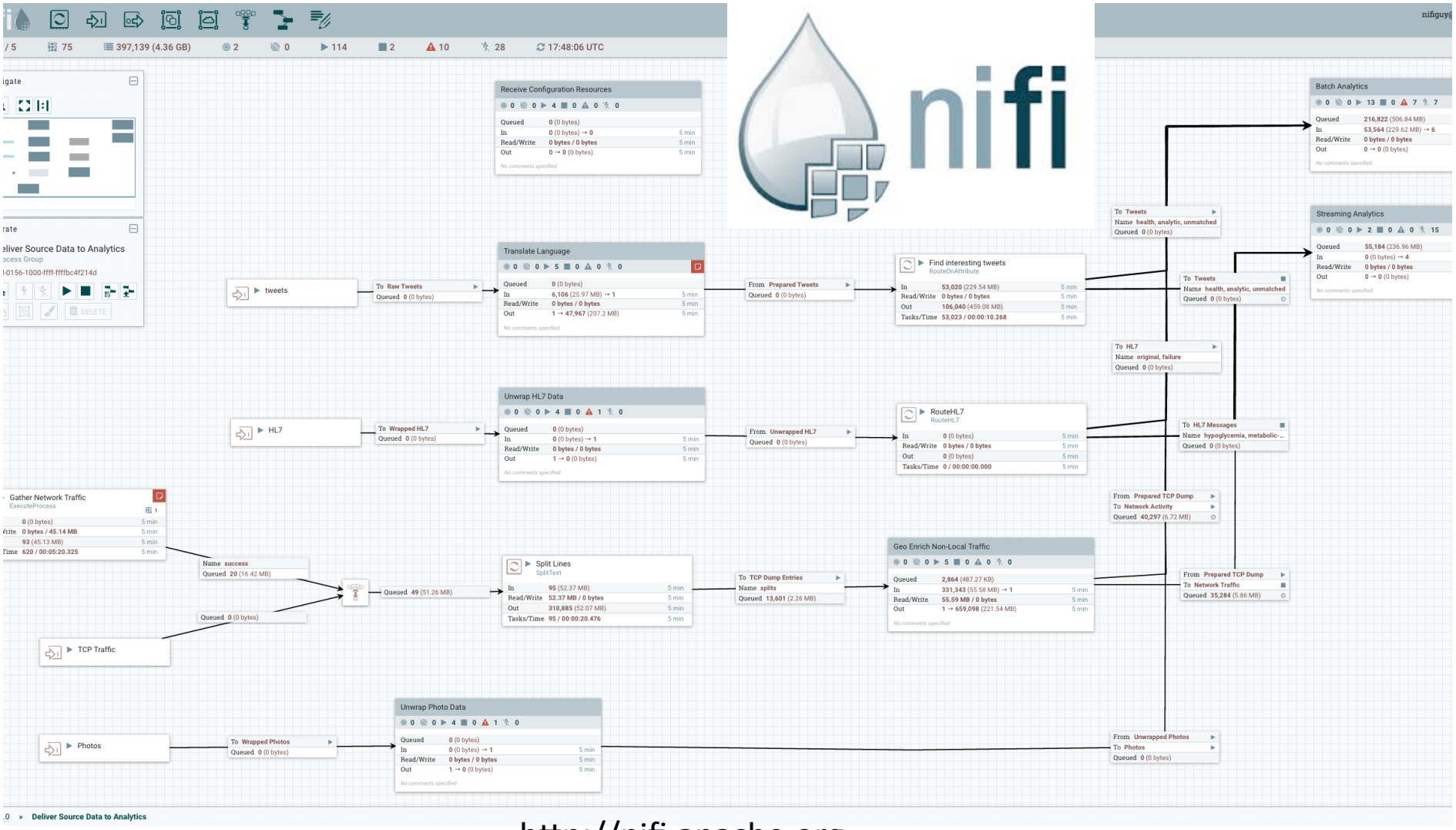


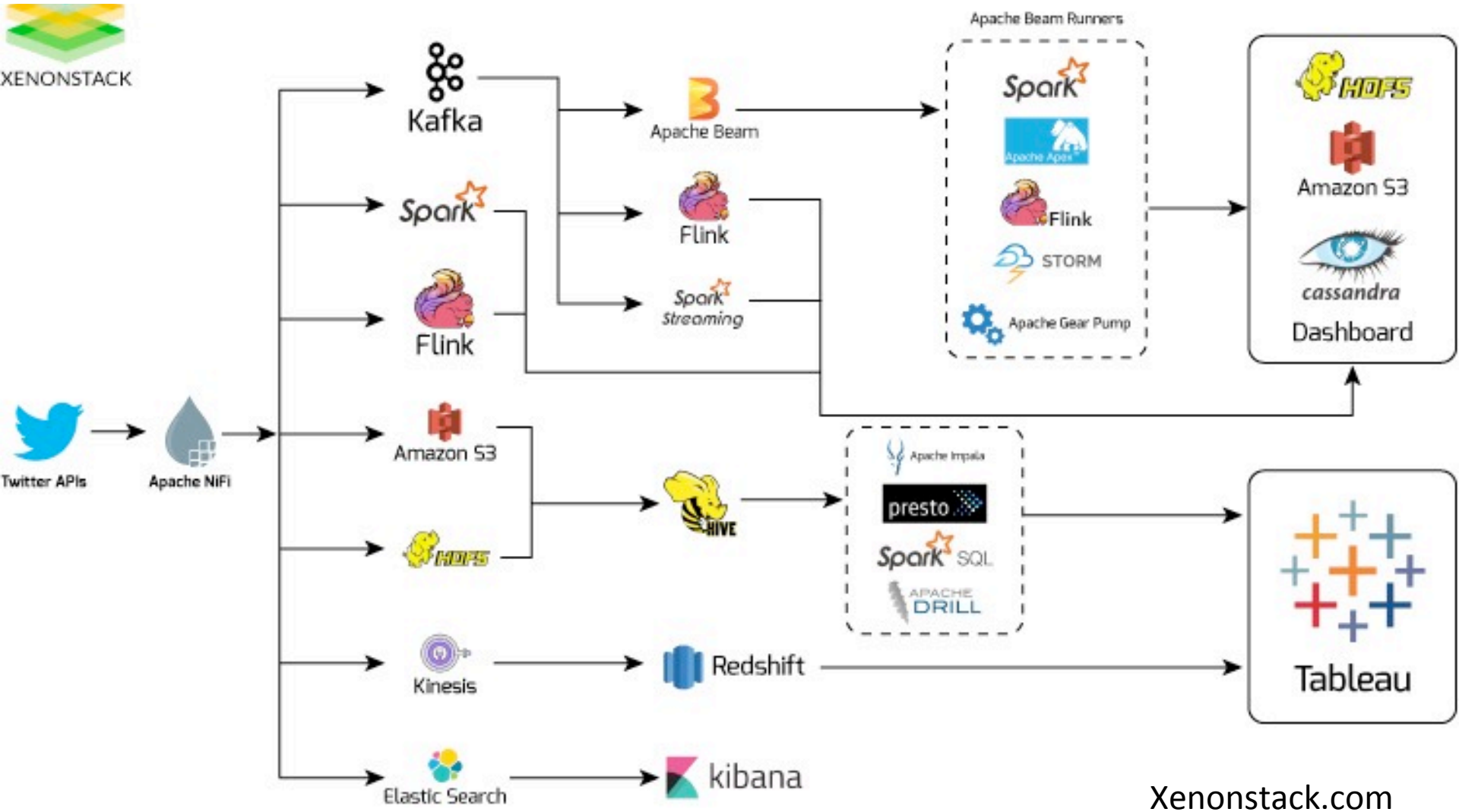
Figure 3.10 Metadata is added to data from internal and external sources.

Data Acquisition and Filtering -NiFi



http://nifi.apache.org

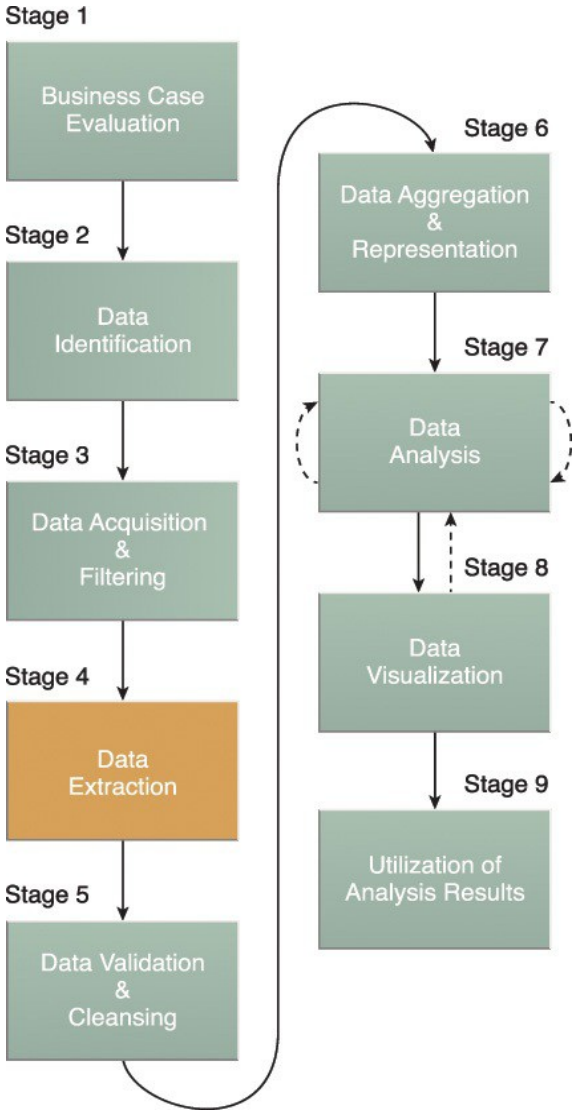
Building a Data Lake with NiFi



Xenonstack.com

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

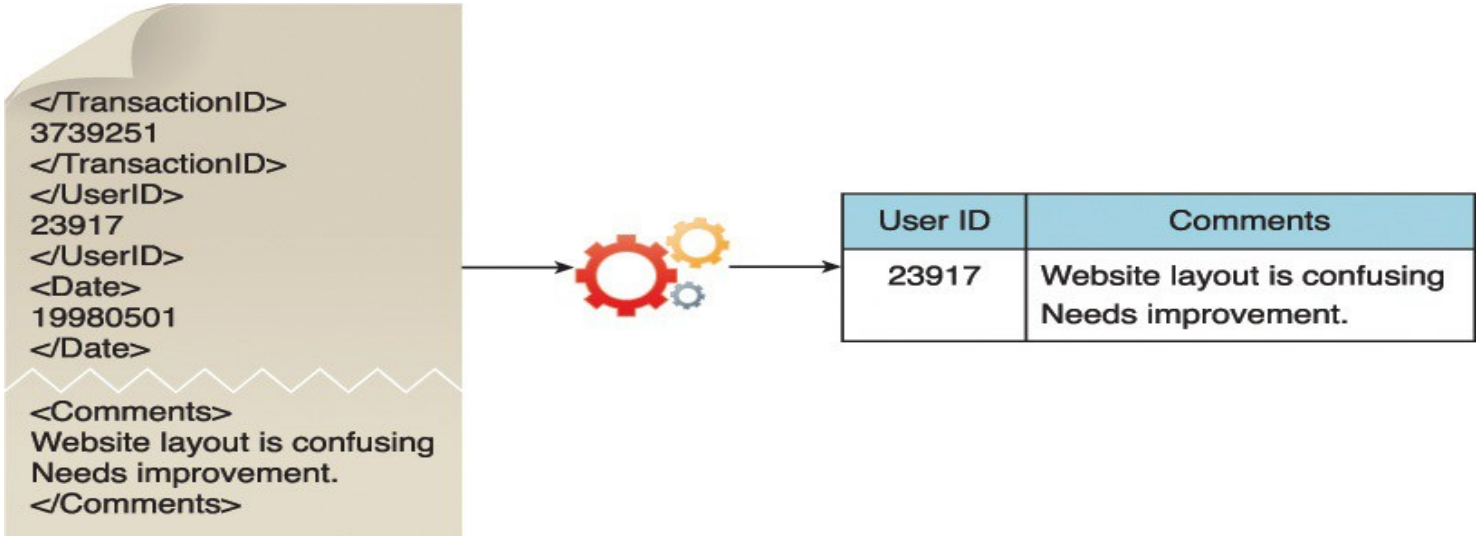


Figure 3.12 illustrates the extraction of comments and a user ID embedded within an XML document without the need for further transformation.

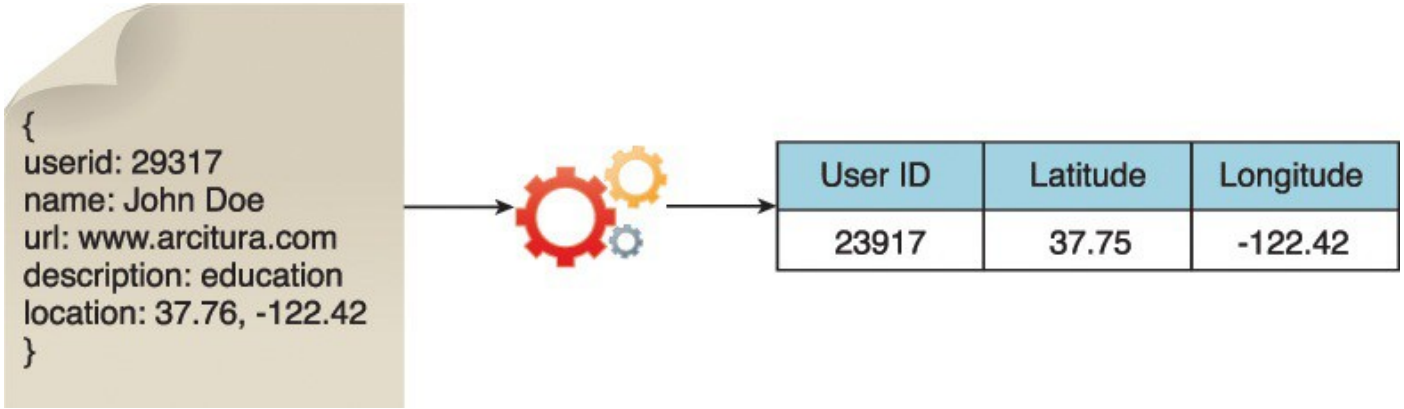
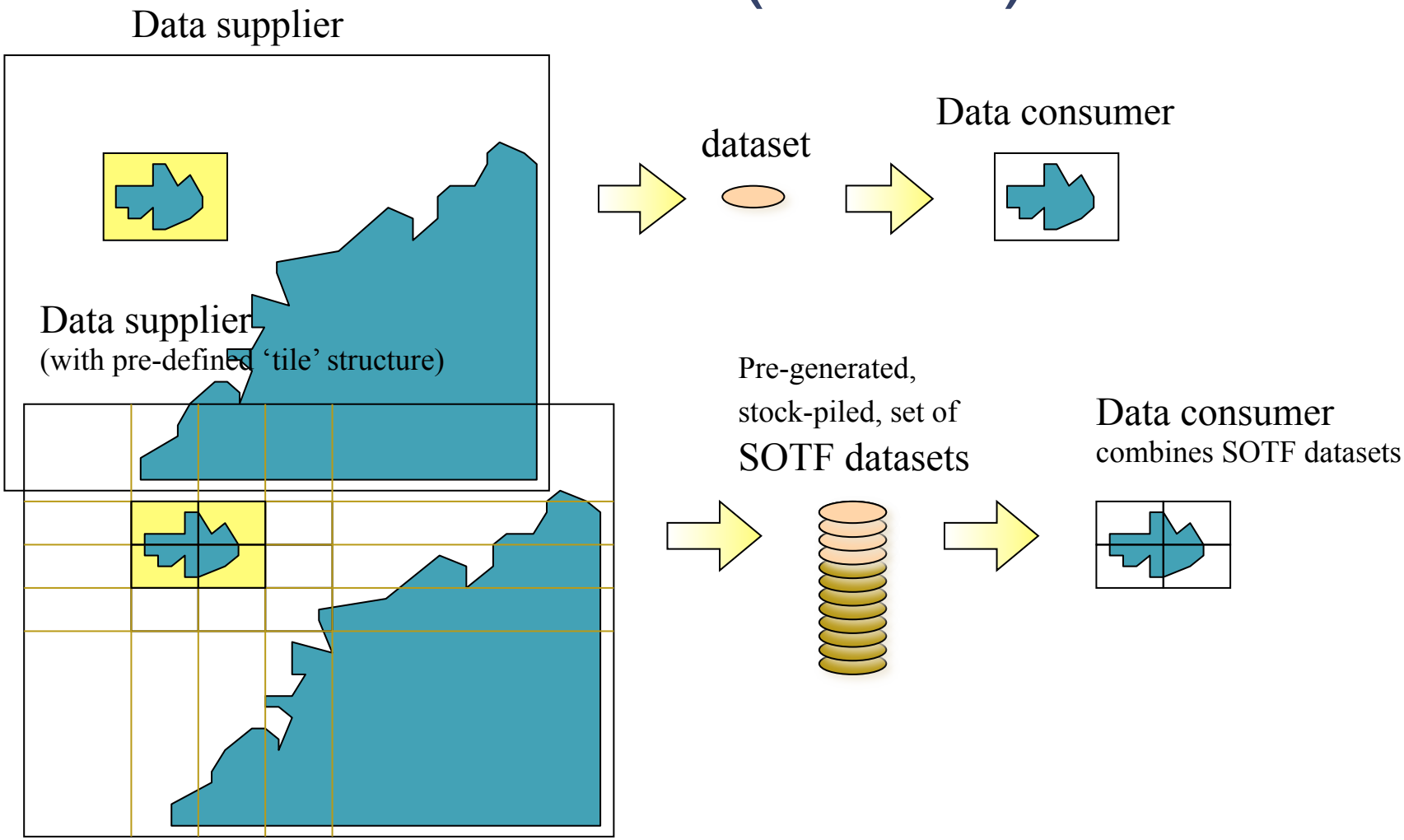


Figure 3.13 demonstrates the extraction of the latitude and longitude coordinates of a user from a single JSON field.

A Prototype Spatial Object Transfer Format (SOTF)

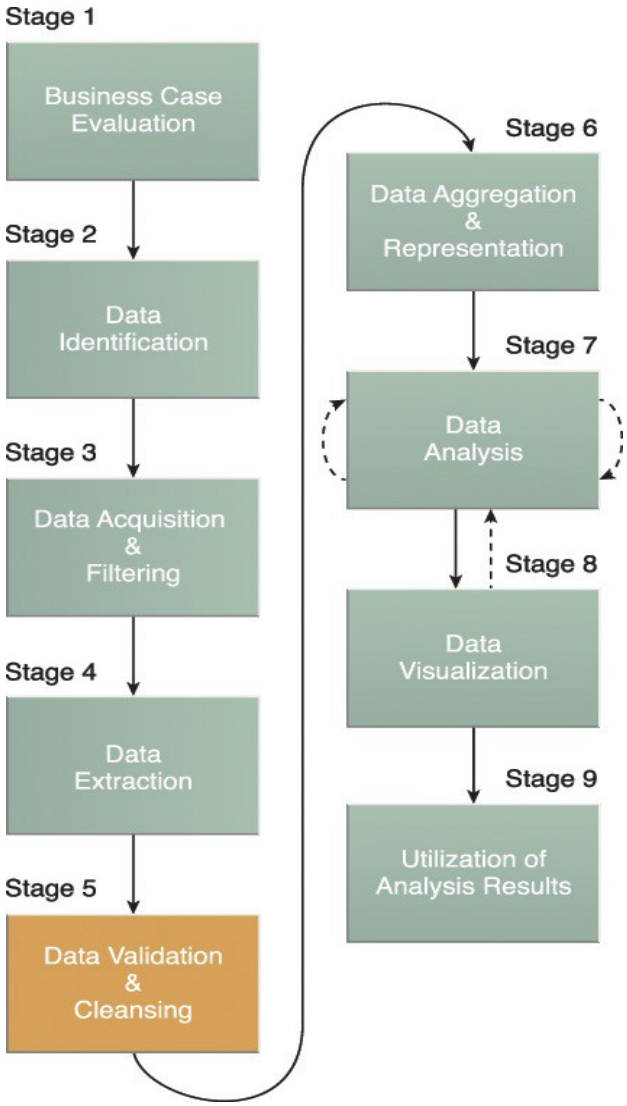
- SOTF supports multiple inheritance of feature types
- SOTF supports light-weight, binary feature relationships
- SOTF it was originally designed to handle complex, structured geospatial data;
  - *it does not support methods and behaviour.*
- SOTF has an object-oriented schema with:
  - features and 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

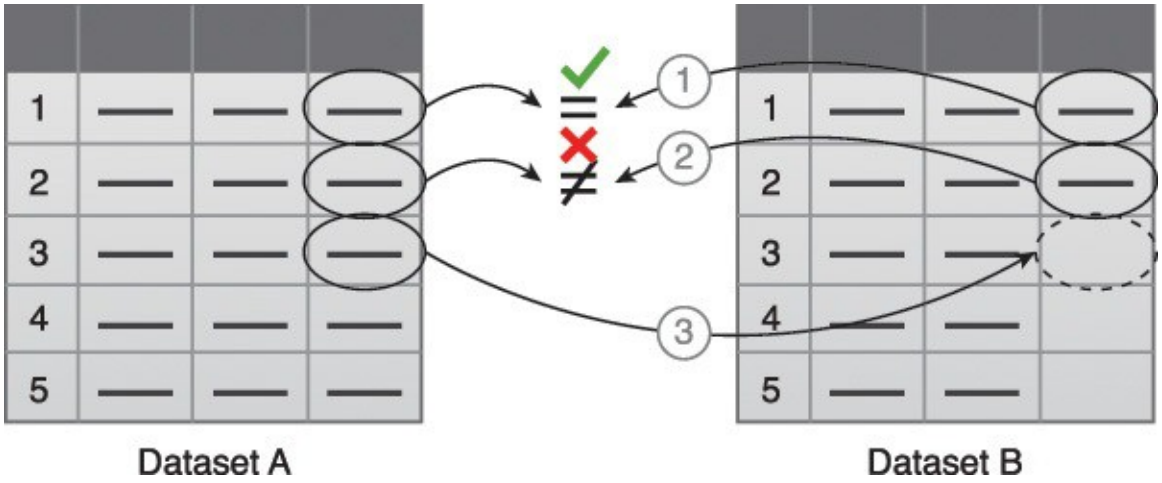


Figure 3.15 Data validation can be used to examine interconnected datasets in order to fill in missing valid data.

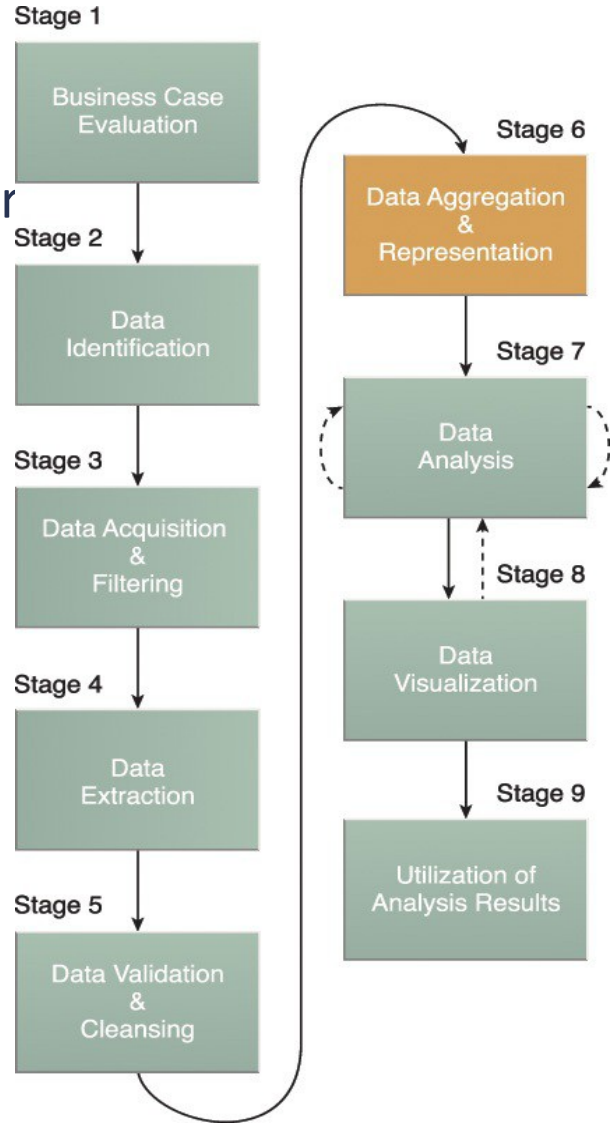
Figure 3.16 The presence of invalid data is resulting in spikes. Although the data appears abnormal, it may be indicative of a new pattern.

Data Aggregation & Representation

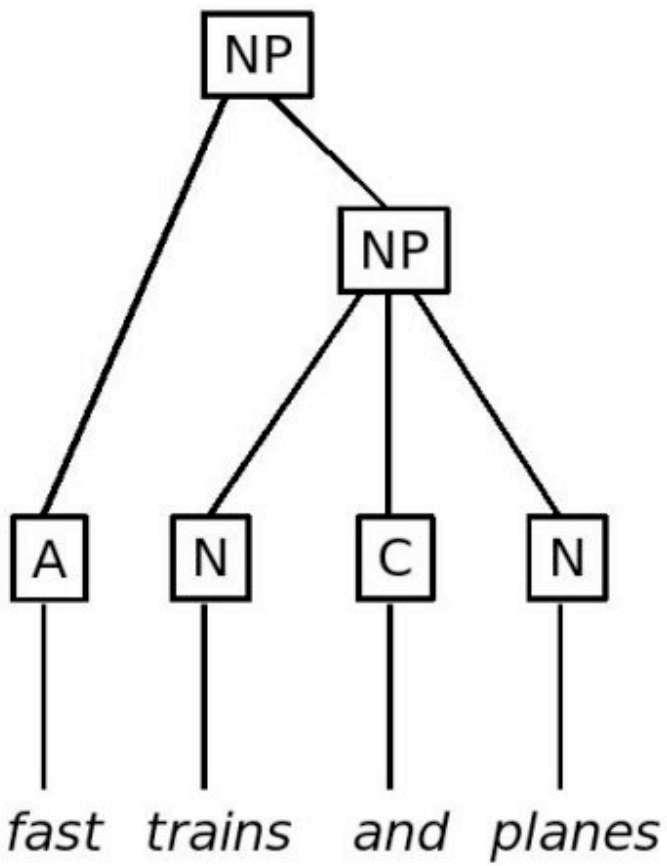
The Data Aggregation and Representation **stage is dedicated to** integrating multiple datasets together to arrive at **a unified view**.

Performing this stage can become complicated because of differences in:

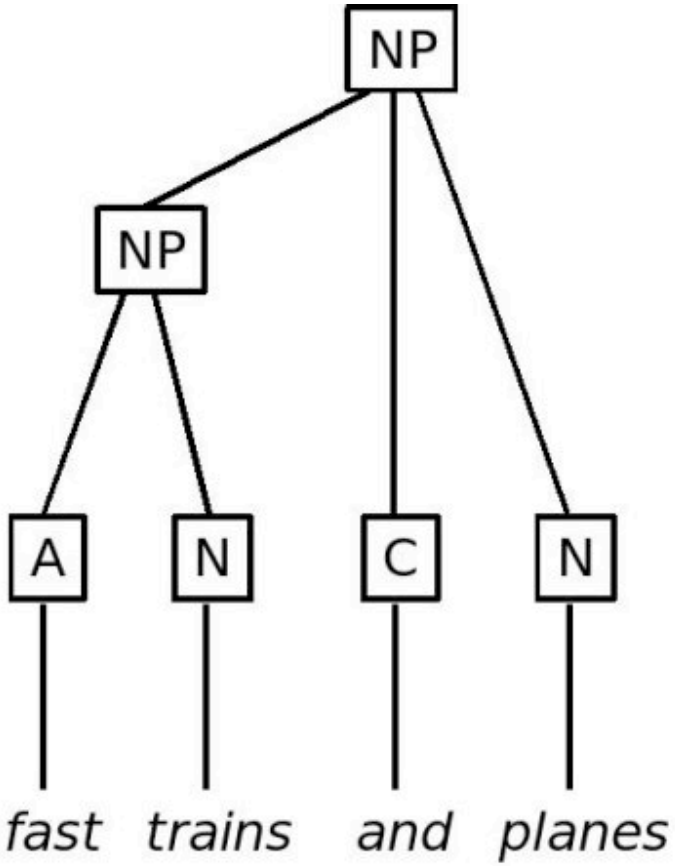
- ***Data Structure*** – Although the data format may be the same, the data model may be different.
- ***Semantics*** – A value that is labeled differently in two different datasets may mean the same thing, for example “surname” and “last name.”



Data Aggregation & Representation



Interpretation A



Interpretation B

Data Aggregation & Representation

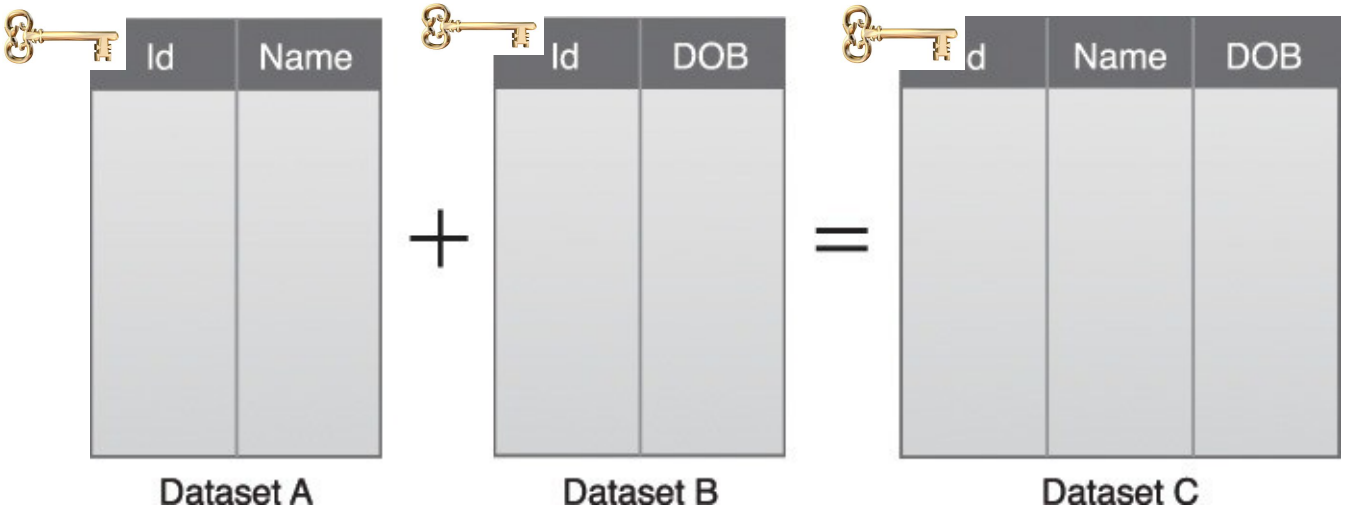


Figure 3.18 A simple example of data aggregation where two datasets are aggregated together using the Id field.

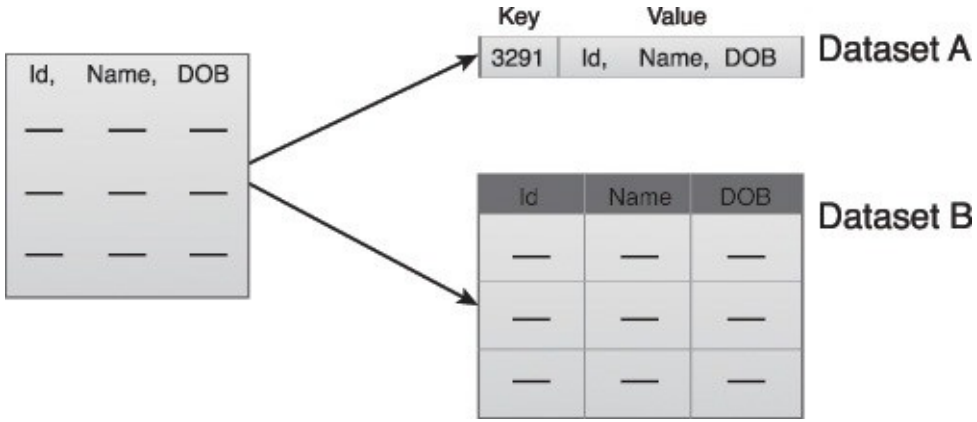


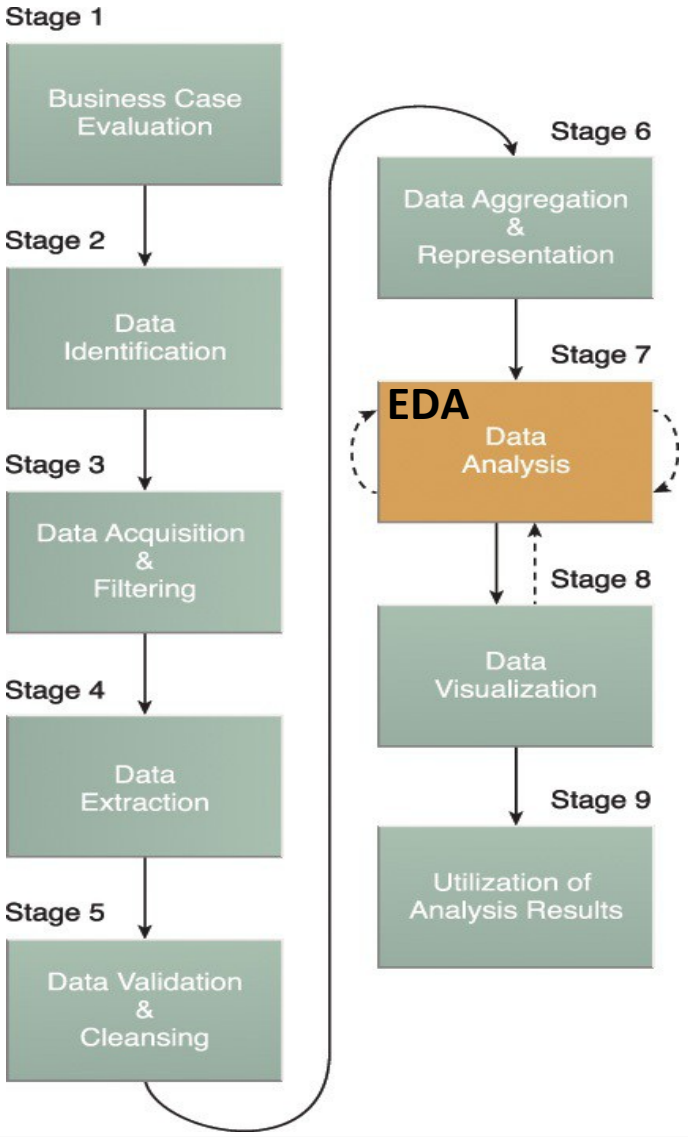
Figure 3.19 Dataset A and B can be combined to create a standardized data structure with a Big Data solution.

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.

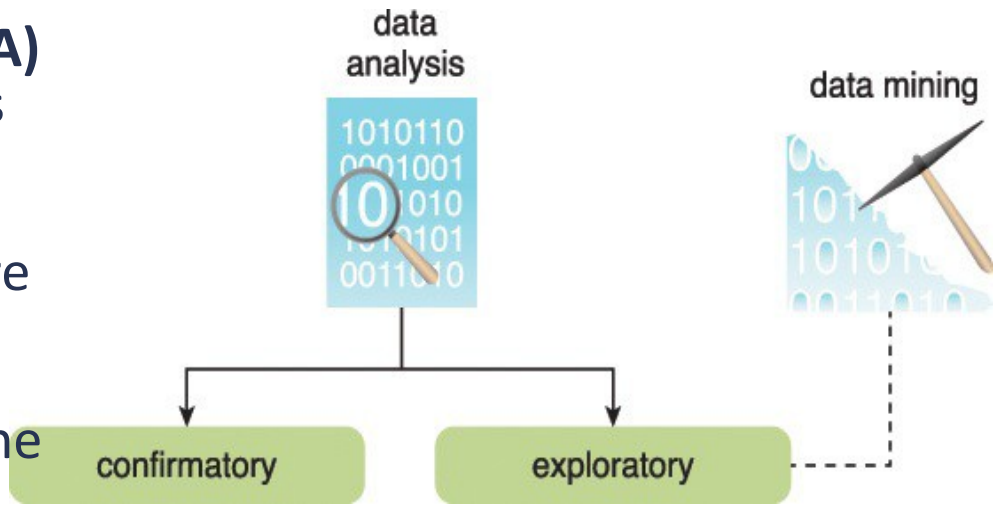
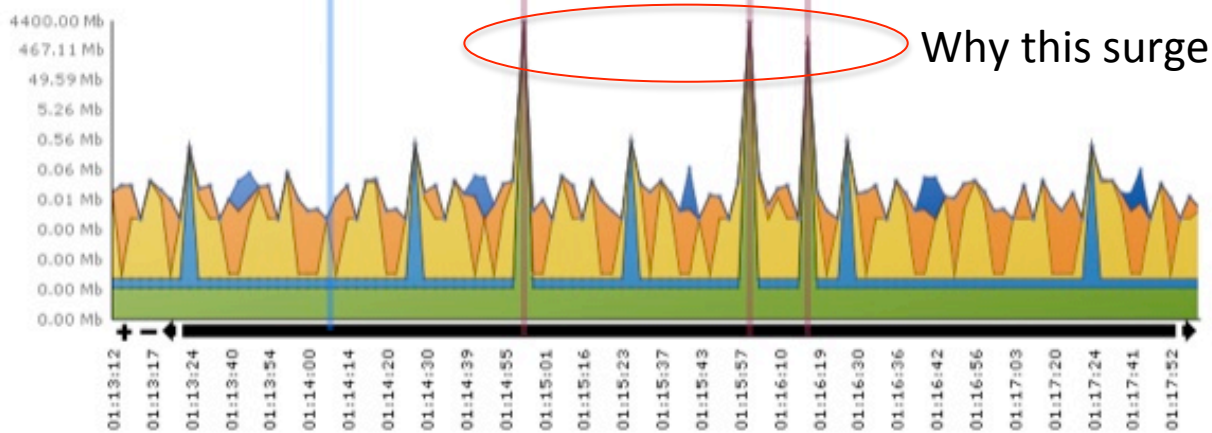


Figure 3.21 Data analysis can be confirmatory or exploratory analysis.

EDA and CDA

Just explore the data to see what is there.

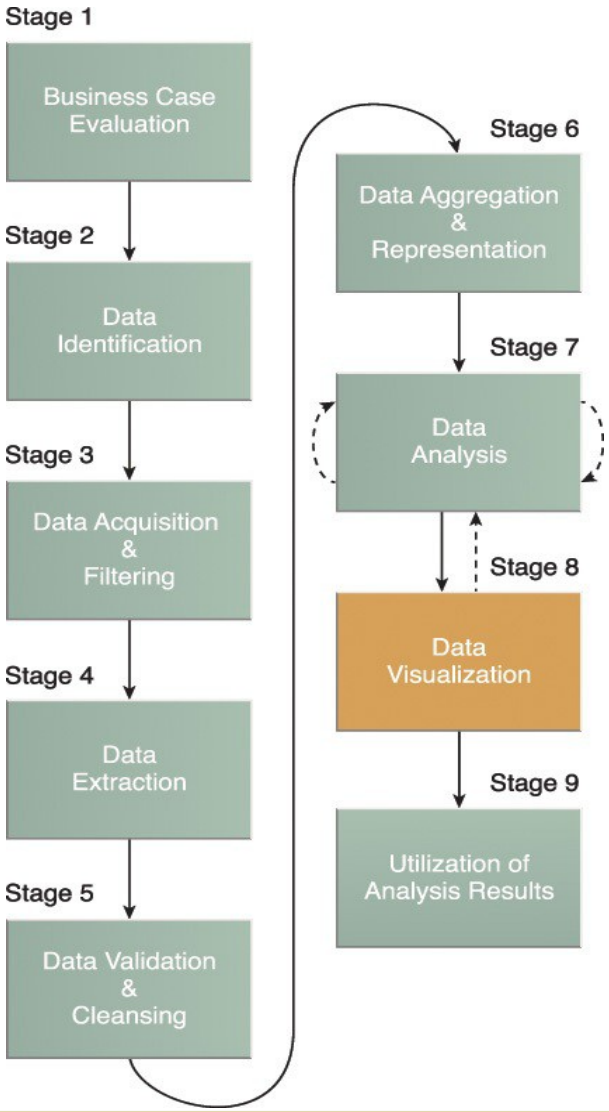


Why this surge in netflow?

Source: sflow.com

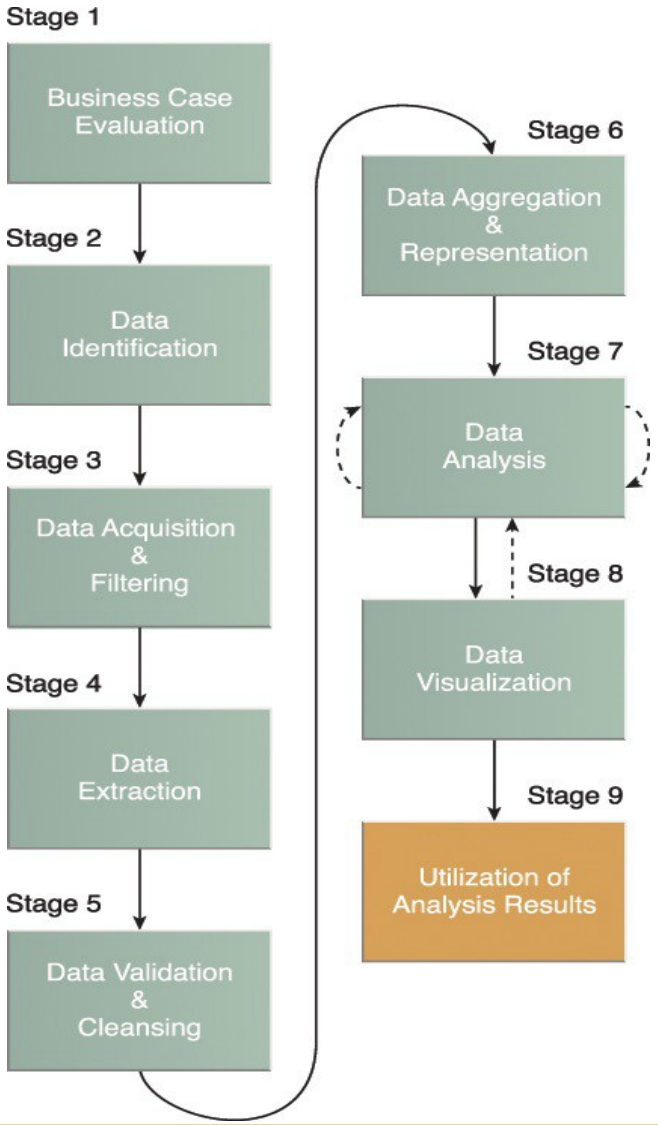
Data Visualization

- 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 d**etermining 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.



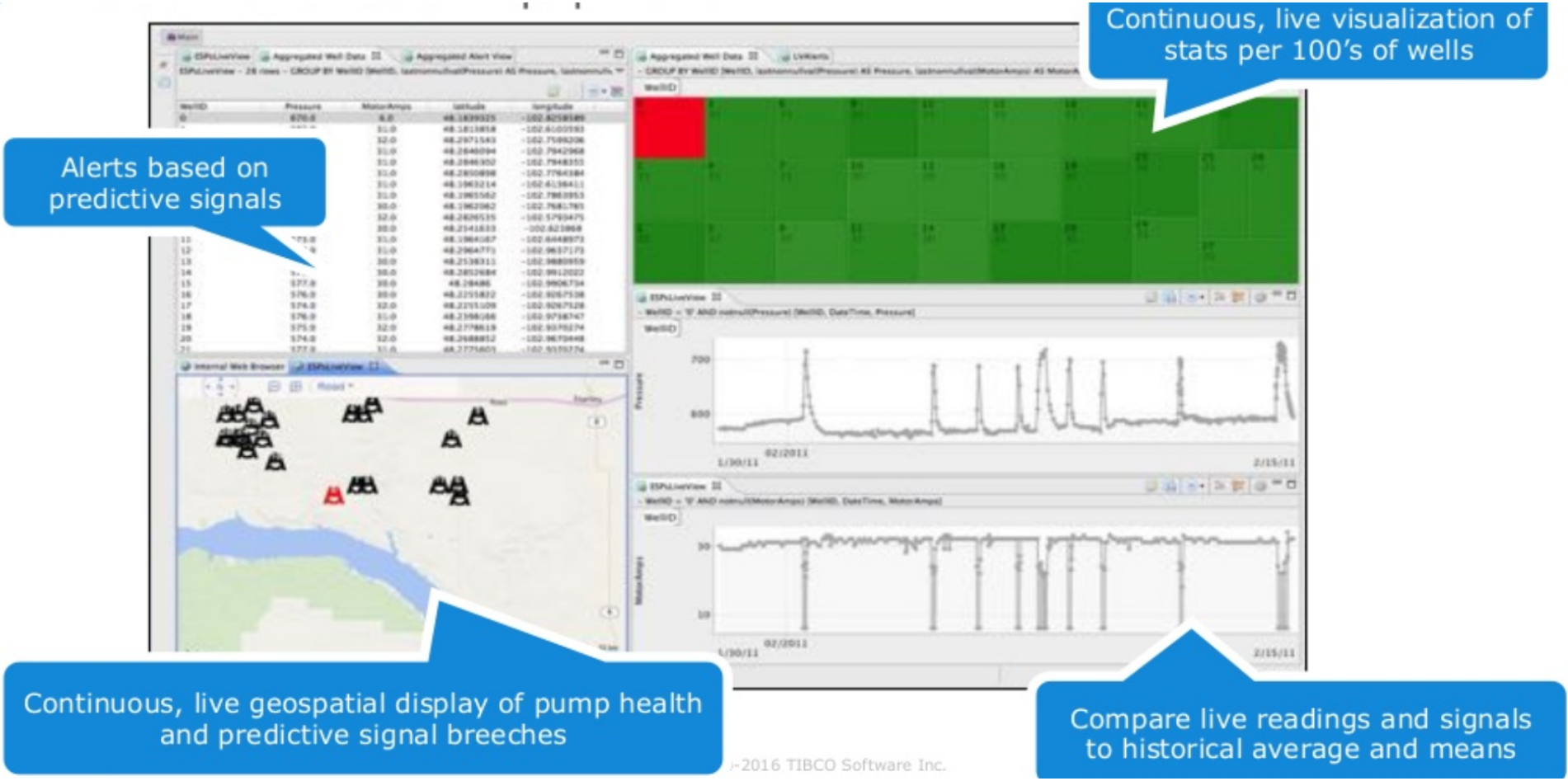
Enterprise Technologies and Big Data Business Intelligence

Data is collected from throughout the enterprise and warehoused in a data warehouse, analytical processing systems answer more complex queries and provide deeper insight into business operations.

It is from these data stores and processing that management gains insight into broader corporate performance and KPIs which include

- Operations/Research Storage and Processing Technologies
  - Stream Processing
  - Batch Processing
- Business Storage and Processing Technologies
  - OLAP (Vertica, Netezza, Hadoop)
  - OLTP-oriented databases: (Postgres, MySQL, VoltDB, Oracle),
- Hybrid OLTP and OLAP: (HBase, Cassandra, MongoDB)
- Extract Transform Load (ETL) Processes
- Data Warehouses
- Data Marts

Why Streaming?



Source: Tibco.com

Streaming and OLTP

Streaming Processing

- Under the streaming model, data is fed into analytics tools piece-by-piece. The processing is usually done in real time.
- Online Transaction Processing (OLTP)
  - OLTP is a software system that processes transaction-oriented data. The term "online transaction" refers to the completion of an activity in real-time and is not batch-processed. OLTP systems store operational data that is normalized. This data is a common source of structured data and serves as input to many analytic processes

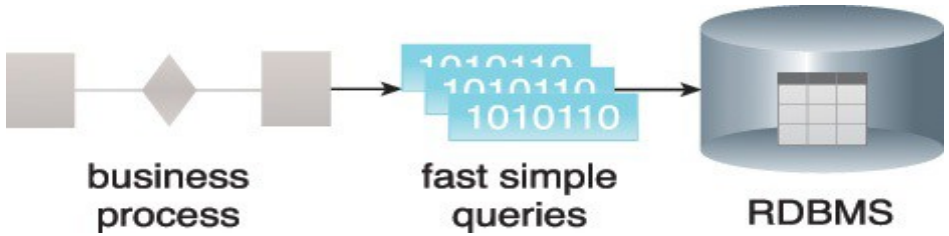


Figure 4.1 OLTP systems perform simple database operations to provide sub-second response times.

Batch and OLAP

Batch Processing

- Under the batch processing model, a set of data is collected over time, then fed into an analytics system. In other words, you collect a batch of information, then send it in for processing.

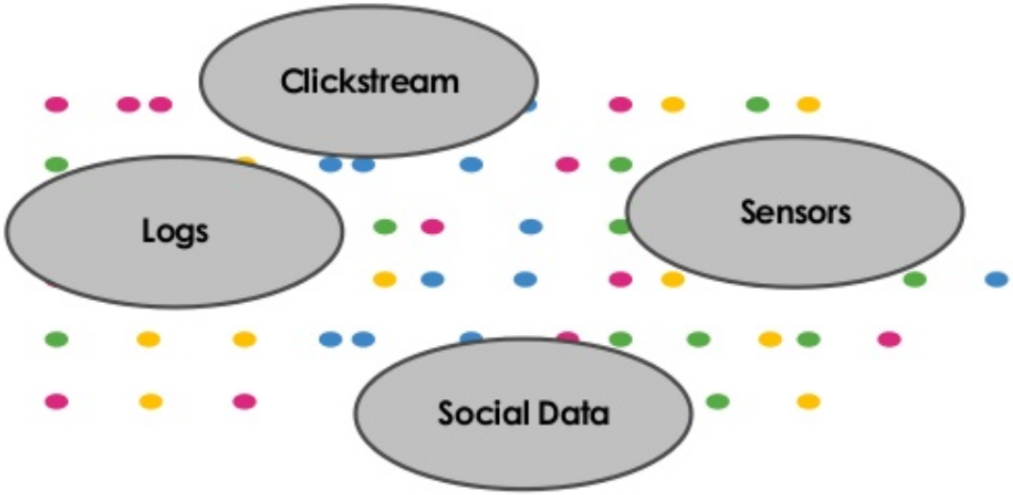
Online Analytic Processing (OLAP)

- Online analytical processing (OLAP) systems are used for processing data analysis queries. OLAPs form an integral part of business intelligence, data mining and machine learning processes. They are relevant to Big Data in that they can serve as both a data source as well as a data sink that is capable of receiving data. They are used in diagnostic, predictive and prescriptive analytics

Figure 4.2 OLAP systems use multidimensional databases.

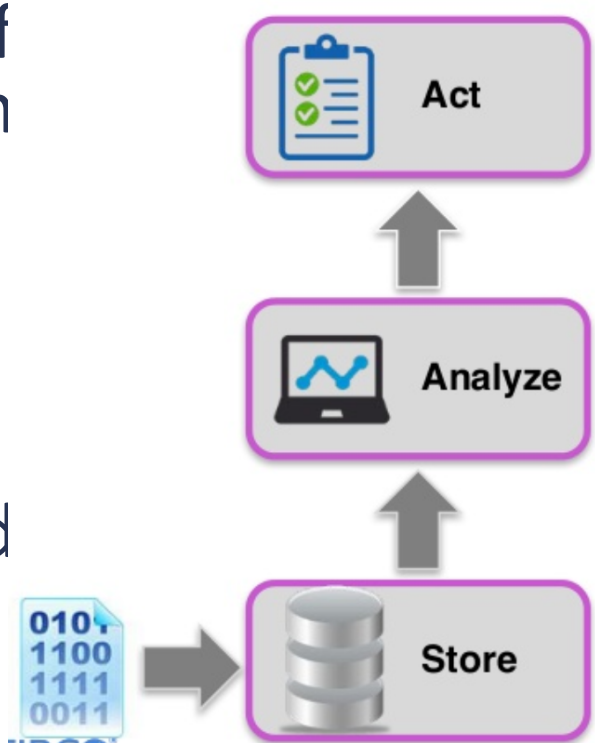
Streaming Analytics: "What Is A Stream"

- Consists of pieces of data typically generated due to a change in state:
  - One or more identifiers
  - Timestamp & payload
  - Immutable: copy then change or delete
- Typically unbounded; there is no end to the data
  - Batch dataset: "bounded"
- Can be raw or derived



Traditional Data Processing – Request & Response

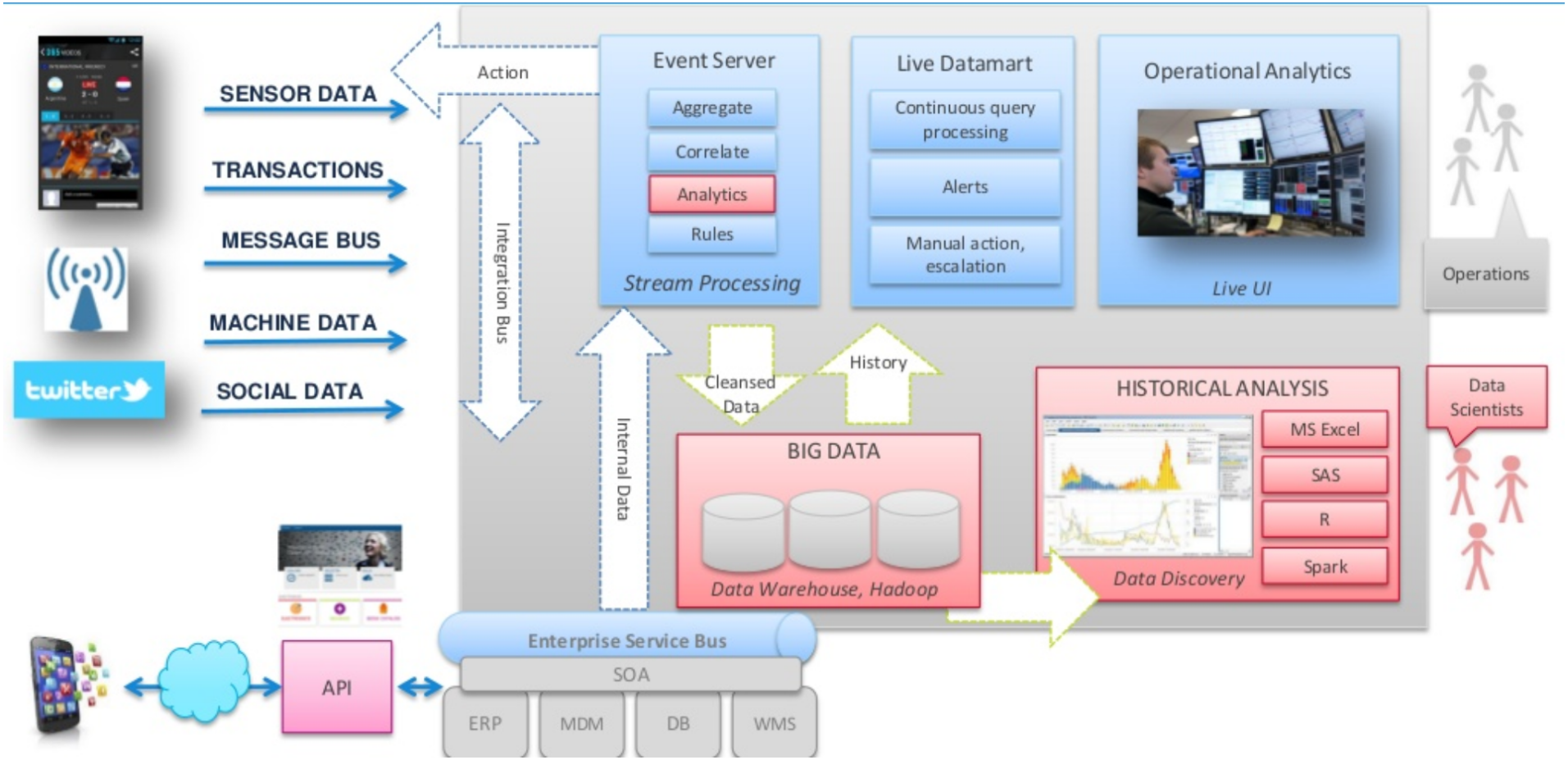
- Data is collected from a variety of sources, and placed in a persisten store
  - Relational database
  - NoSQL store
  - Hadoop environment
- Analytical processes are executed against the stored data to detect opportunities or threats
- Actions are identified, delivered, and executed across various business channels



The New Era: Streaming Analytics

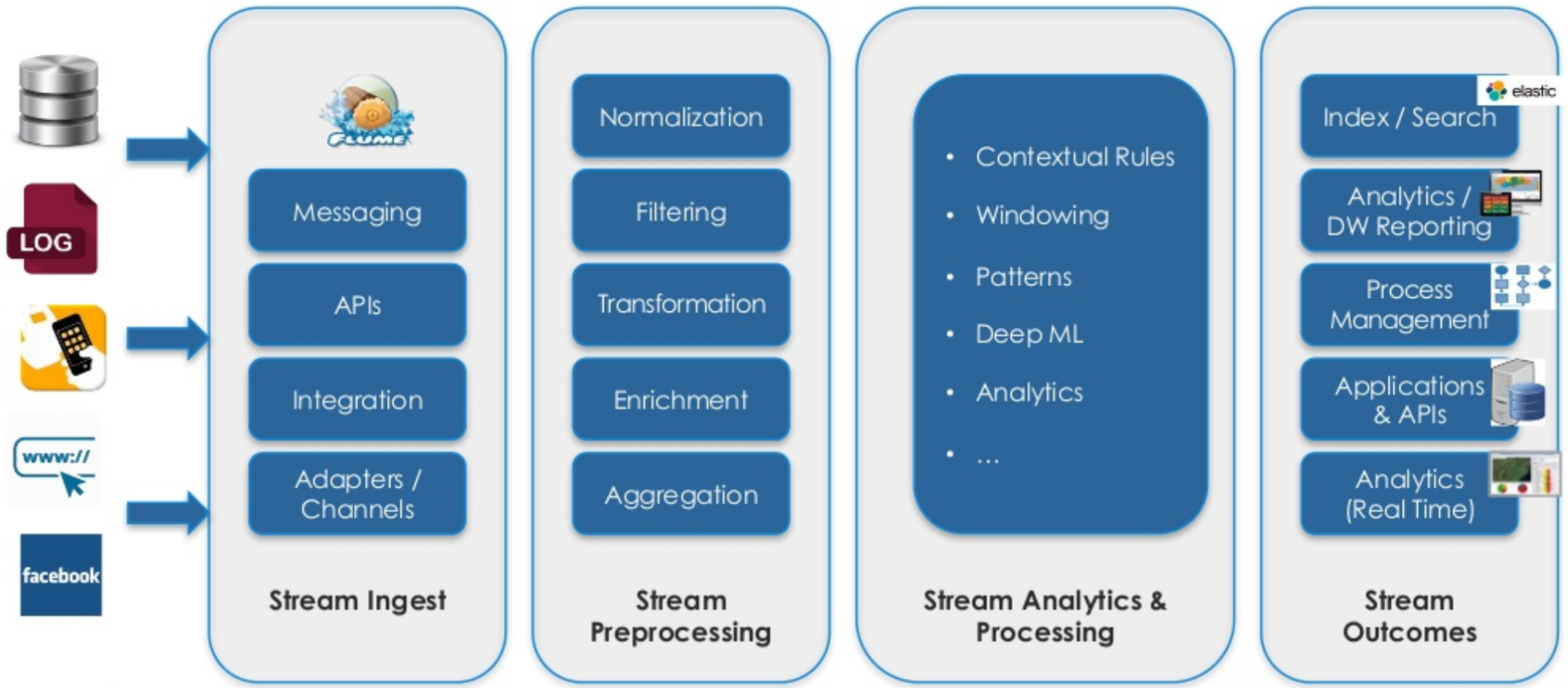
- Events are analyzed and processed in real-time as they arrive.
- Decisions are timely, contextual, and based on fresh data.
- Decision latency is eliminated, resulting in:
  - Superior Customer Experience,
  - Operational Excellence,
  - Instant Awareness and Timely Decisions

Streaming Analytics Reference Architecture



Source: tibco.com

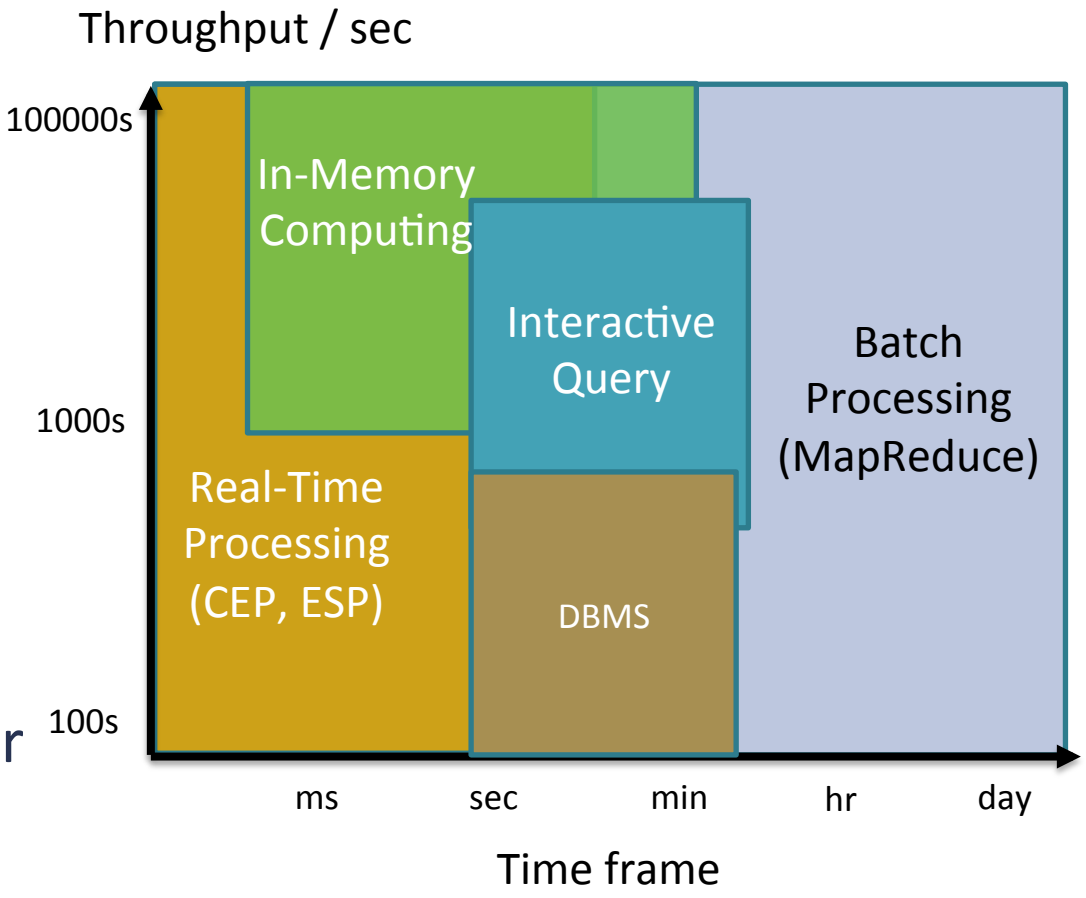
Streaming Analytics Processing Pipeline



Source: tibco.com

Real-time Batch Comparison

- Apache Spark processes data in-memory while Hadoop MapReduce persists back to the disk after a map or reduce action, so Spark should outperform Hadoop MapReduce.
- Spark needs a lot of memory. Much like standard DBs, it loads a process into memory and keeps it there until further notice, for the sake of caching.



Extract Transform Load (ETL)

- Extract Transform Load (ETL) is a process of loading data from a source system into a target system. The source system can be a database, a flat file, or an application. Similarly, the target system can be a database or some other storage system.

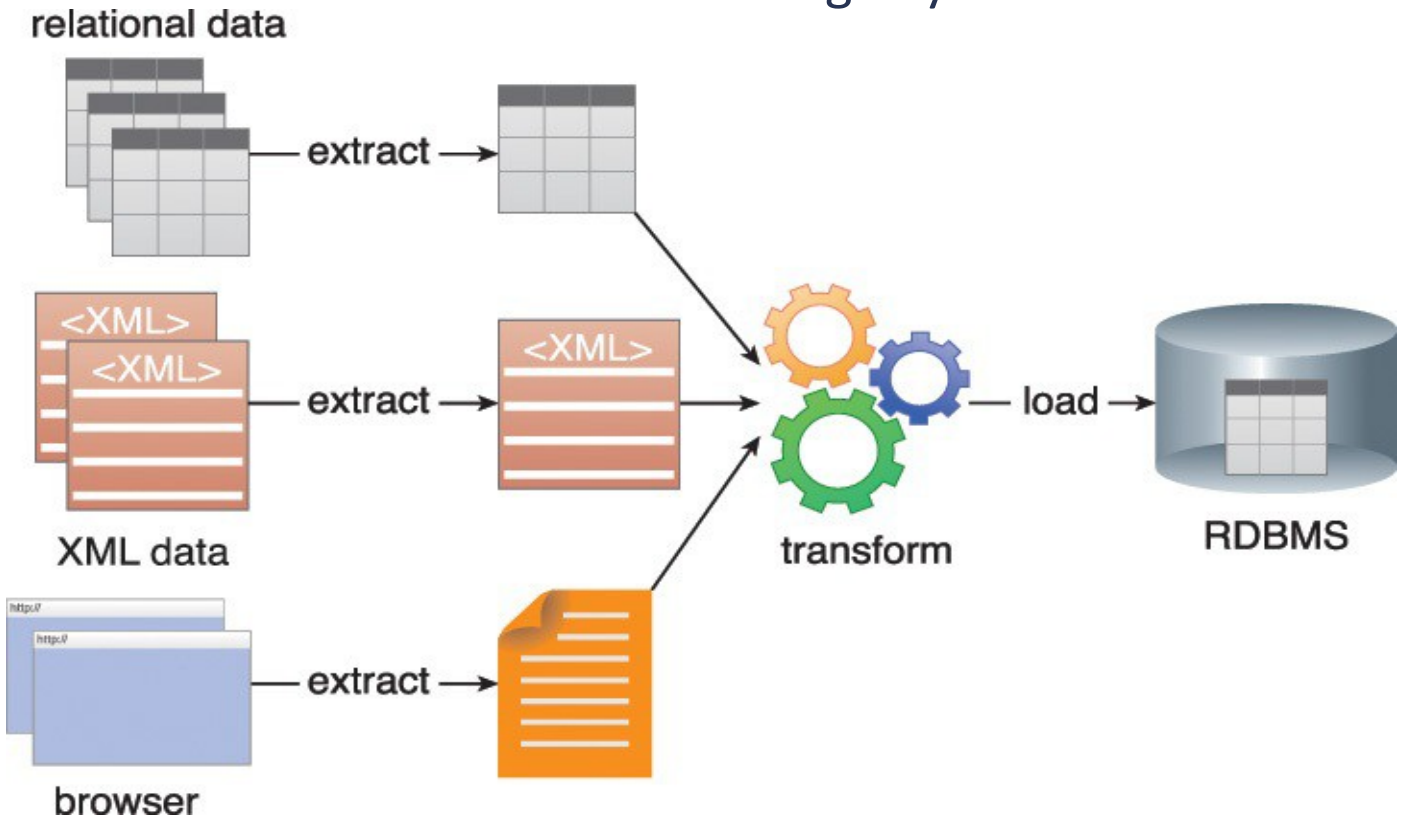


Figure 4.3 An ETL process can extract data from multiple sources and transform it for loading into a single target system.

Data Warehouse

A data warehouse is a central, enterprise-wide repository consisting of historical and current data. Data warehouses are heavily used by BI to run various analytical queries, and they usually interface with an OLAP system to support multi-dimensional analytical queries.

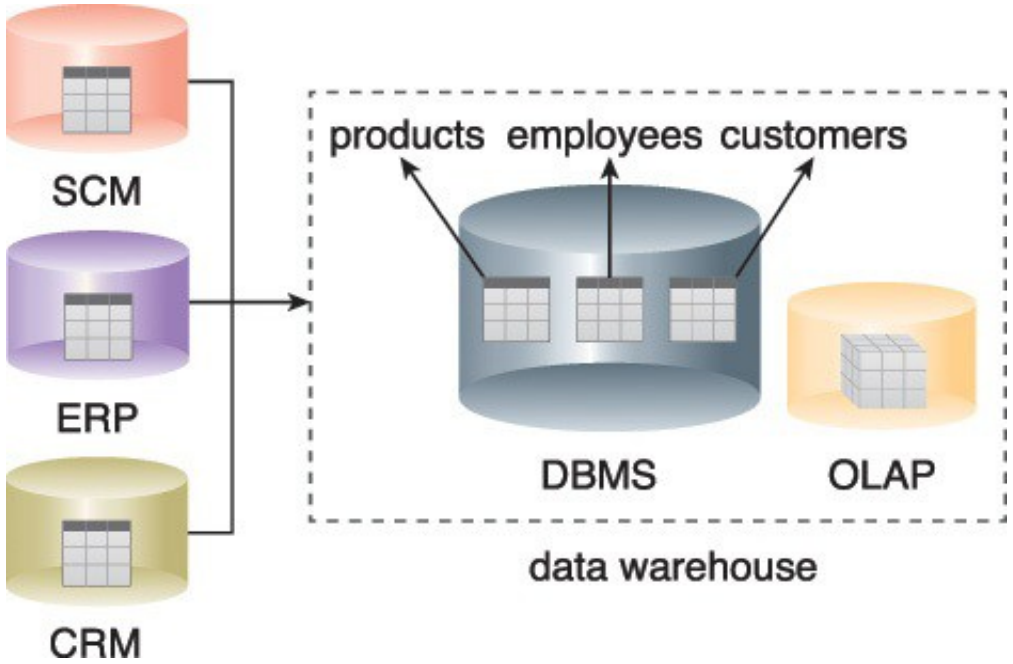


Figure 4.4 Batch jobs periodically load data into a data warehouse from operational systems like ERP, CRM and SCM.

Data Marts

A data mart is **a particular subset of the data** stored in a data warehouse that typically belongs to a department, division, or specific line of business.

Enterprise-wide data is collected and business entities are then extracted. Domain specific entities are persisted into the data warehouse via an ETL process.

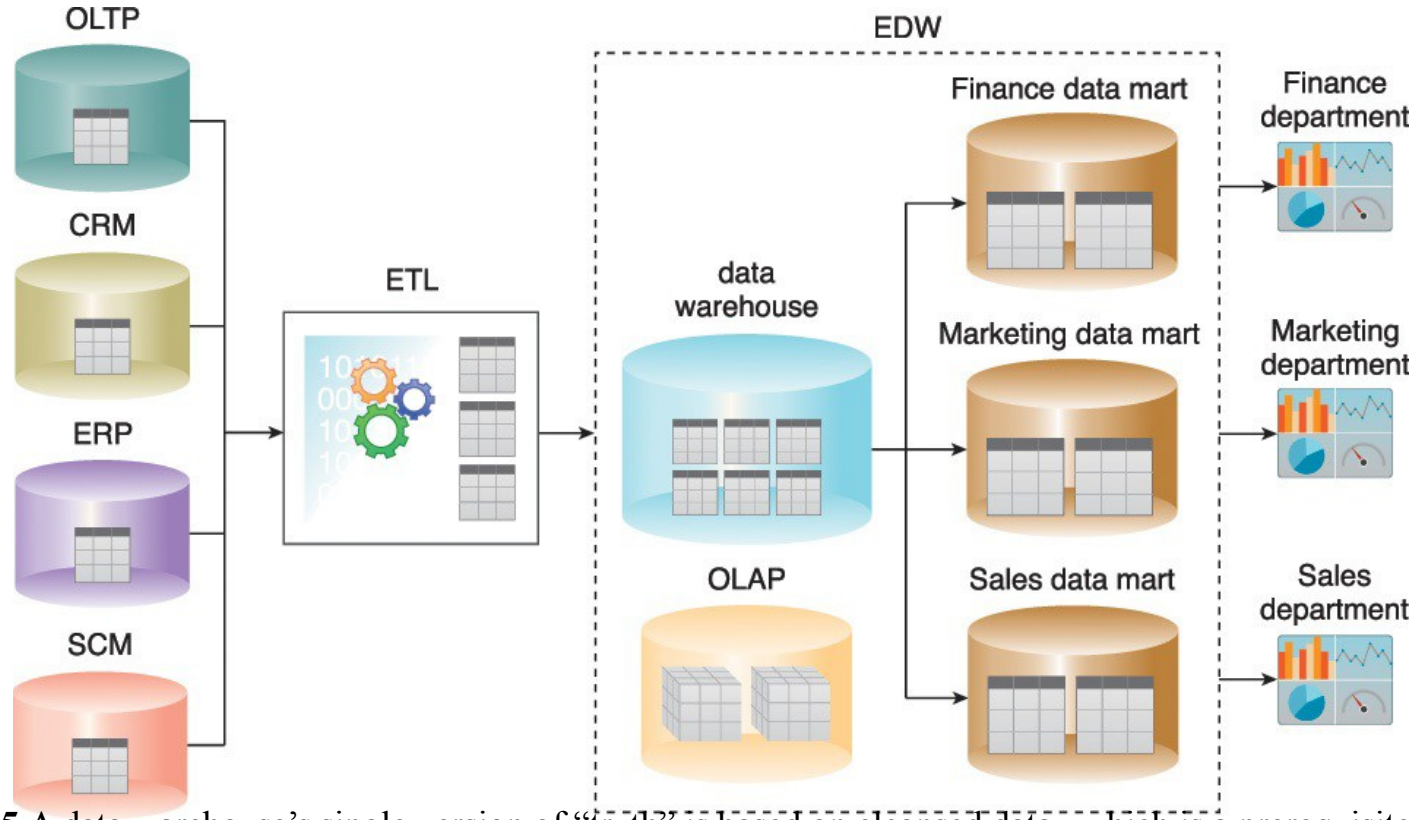


Figure 4.5 A data warehouse's single version of "truth" is based on cleansed data, which is a prerequisite for accurate and error-free reports, as per the output shown on the right.

Ad-hoc Reports and Dashboards

Traditional BI primarily utilizes **descriptive** and **diagnostic** analytics to provide information on historical and current events. It is not "intelligent" because it only provides answers to correctly formulated questions through:

- **Ad-hoc reporting** is a process that involves manually processing data to produce custom-made reports
- **Dashboards** provide a holistic view of key business areas. The information displayed on dashboards is generated at periodic intervals in realtime or near-realtime.

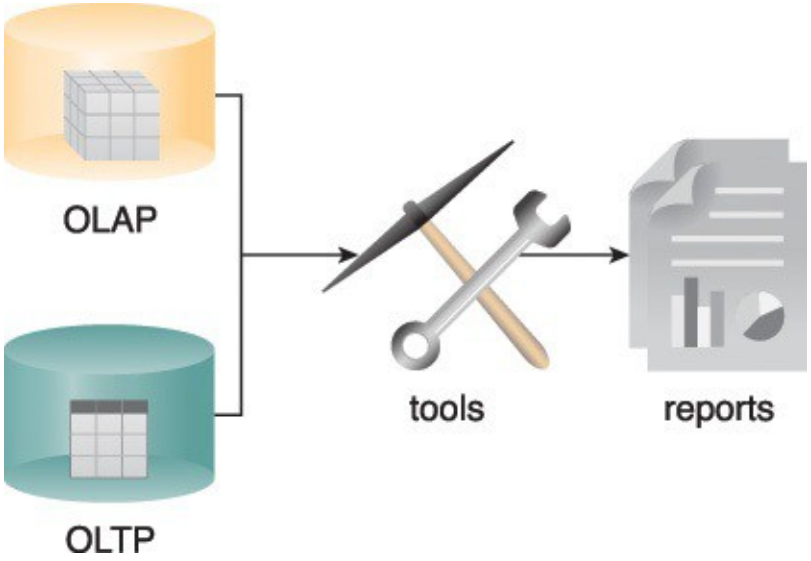


Figure 4.6 OLAP and OLTP data sources can be used by BI tools for both ad-hoc reporting and dashboards.

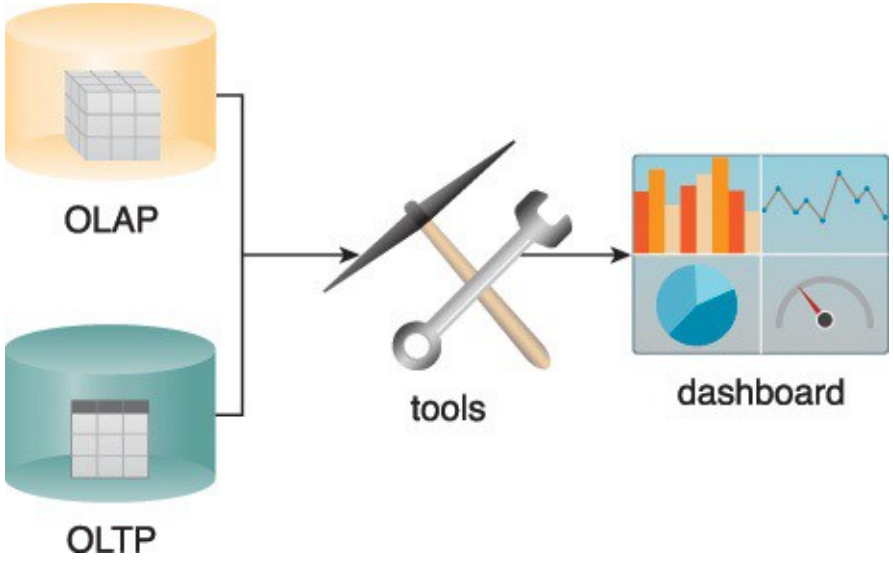


Figure 4.7 BI tools use both OLAP and OLTP to display the information on dashboards.

Traditional Business Intelligence

Traditional BI uses data warehouses and data marts for reporting and data analysis because they allow complex data analysis queries with multiple joins and aggregations to be issued

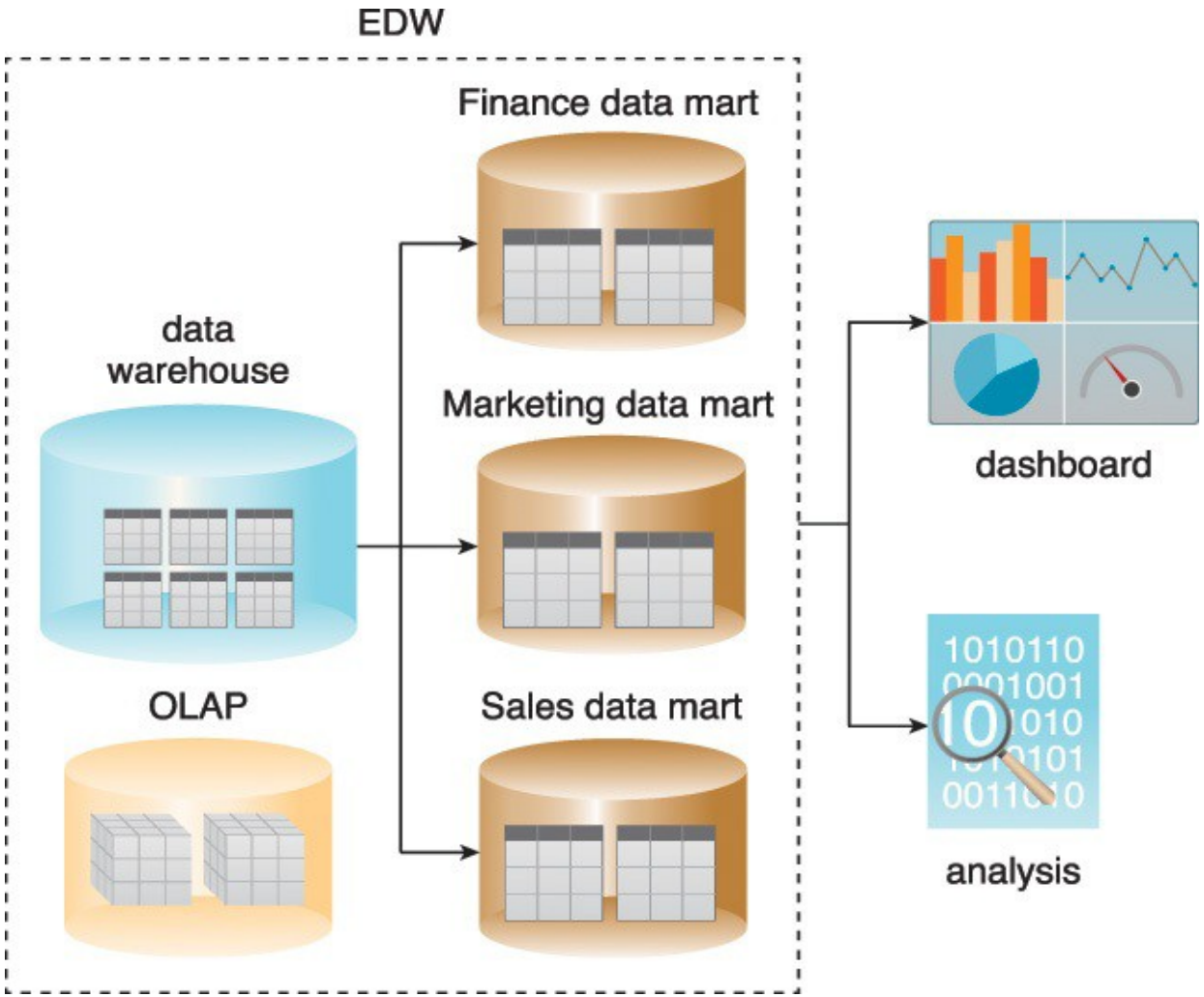


Figure 4.8 An example of traditional BI.

Big Data Business Intelligence

- Big Data BI builds upon traditional BI by acting on the cleansed, consolidated enterprise- wide data in the data warehouse and combining it with semi-structured and unstructured data sources.
- It comprises both predictive and prescriptive analytics to facilitate the development of an enterprise-wide understanding of business performance.

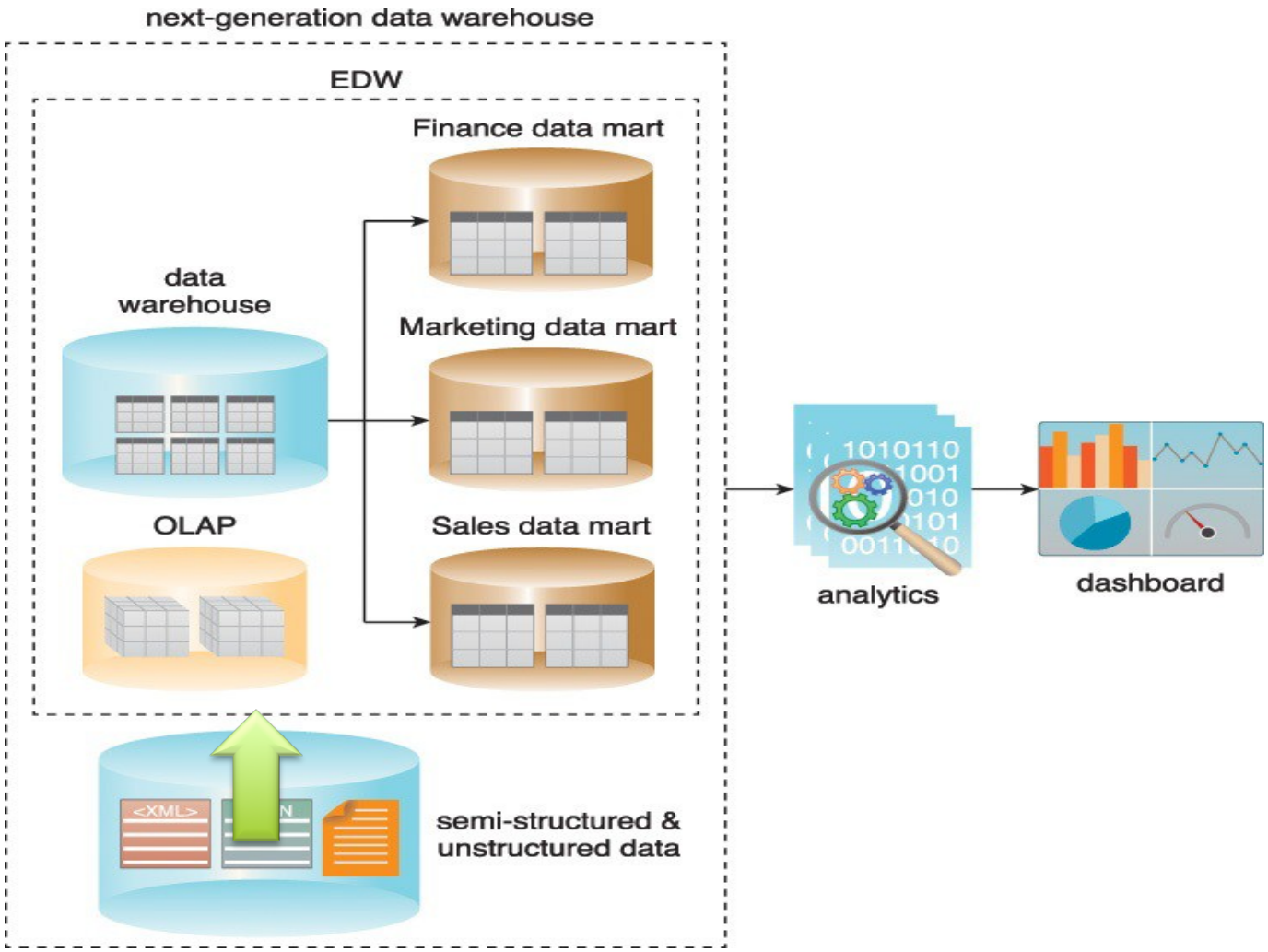
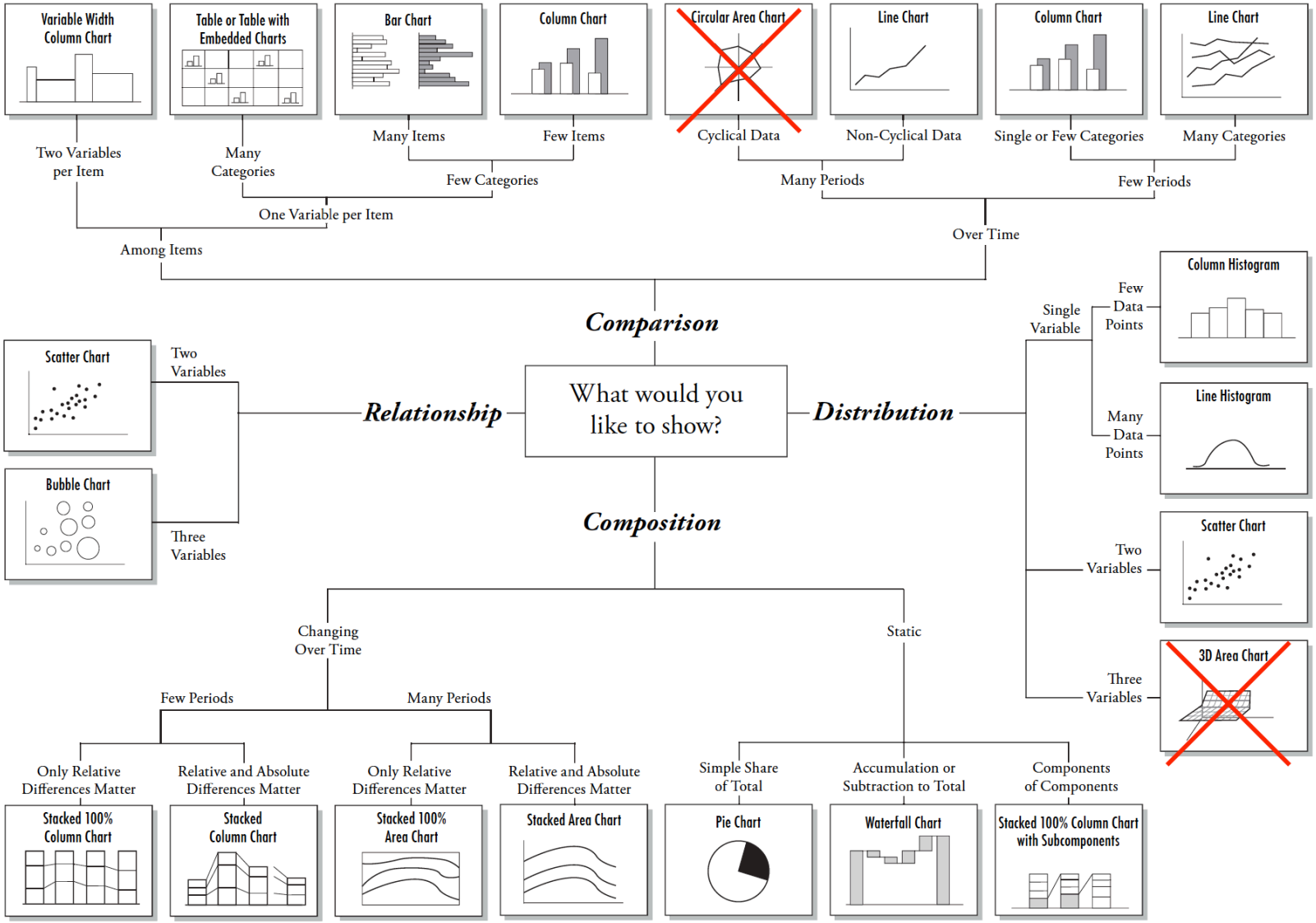


Figure 4.9 A next-generation data warehouse.

Traditional & Big Data Visualizations

- Data visualization is a technique whereby analytical results are graphically communicated, interpreted, and portrayed using elements like **charts, maps, data grids, infographics and alerts.**
- Graphically representing data can make it easier to understand reports, view trends and identify patterns.
- Big Data solutions require data visualization tools that can **seamlessly connect to structured, semi-structured and unstructured data sources** and are further capable of handling millions of data records.
- Data visualization tools for Big Data solutions generally use **in-memory analytical technologies** that reduce the latency normally attributed to traditional, disk-based data visualization tools.

Visualization Types



Python

- Python is a general-purposed high-level programming language
  - Web development
  - Networking
  - Scientific computing
  - Data analytics
  - …

Python for Data Analytics

- The nature of Python makes it perfect-fit for data analytics
  - Easy to learn
  - Readable
  - Scalable
  - Extensive set of libraries
  - Easy integration with other apps
  - Active community & ecosystem

Portfolio's

- iPython is a Python command shell for interactive computing
- Jupyter Notebook (the former iPython Notebook) is a web-based interactive data analysis environment that supports iPython