Chapter 3

Instructor: Houssein Dhayne
houssein.dhayne@net.usj.edu.lb

Strategic Considerations and Planning for Successful Big Data Adoption

Overview

- **Big Data Initiatives** are strategic and business-driven, with transformative and innovative potential.
- Transformation activities focus on efficiency, while innovation alters business structures.
- Chapter 3 delves into the planning and adoption of Big Data, highlighting the importance of careful innovation management.

Table of content

Example

A company decides to use Big Data to improve its operations.

- Instead of just trying to do things a little better (like making small changes to existing processes), they decide to completely transform the way they work.
- For instance, if they were a retail company, instead of just improving their current stores, they might use Big Data to completely rethink how they sell products, maybe by creating a personalized online shopping experience.
  - Transformation involves making changes that are low-risk but improve efficiency.
- On the other hand, innovation requires a more significant shift in mindset because it can fundamentally change your business. For instance, it might lead to new products or services.

Challenges

- However, **innovating with Big Data comes with challenges**.
- You need to carefully manage the process.
  - Too much control can stifle creativity,
  - while too little oversight can turn your project into a failed experiment.
- This is where this chapter comes in, it addresses the considerations and planning needed for Big Data adoption.

Planning Considerations

- Big Data adoption involves **various considerations**, such as security, privacy, governance, and data management.
  - New governance processes and decision frameworks are essential to manage Big Data's nature and implications.
- Big Data **Analytics Lifecycle**
  - Introduction of a lifecycle from establishing a business case to deploying analytic results.
  - Involves stages like identifying, procuring, filtering, extracting, cleansing, and aggregating data.

6

Organization Prerequisites for Big Data Adoption

Big Data frameworks require proper data management and governance.

- Processes, skillsets, and data quality assessment are necessary.
- Planning for the longevity of the Big Data environment is essential for sustained success.
- Example: let's imagine you're **the head of a healthcare data team**, and you want to use Big Data to improve patient care.
  - **Not a Magic Fix:** Big Data frameworks are not magic fixes. You can't just plug them in and expect miracles.
  - **Need a Plan:** To help in healthcare, you need a plan, a set of rules and processes (like a guidebook) for handling and managing the data.
  - **Skills Matter:** your team needs skills to implement and use Big Data effectively.
  - **Check the Data Quality:** You need to check that the data going in is accurate and up-to-date; otherwise, you'll get poor results.
  - **Think Long-Term:** Plan for the future with Big Data. Create a roadmap, like a schedule, to make sure your data system can grow and adapt along with your healthcare needs.

7

Data Procurement -1-

For instance, in the lifecycle of Big Data analytics on the political orientations of society, you might consider procuring various types of external data to gain comprehensive insights. Here are some potential sources of external data:

- **Social Media Data**: Analyzing sentiments, discussions, and trends on platforms like Twitter, Facebook, or Instagram can provide real-time insights into political opinions.
- **News and Media Feeds**: Extracting data from news articles, blogs, and online news sources can help understand public reactions, political events, and emerging trends.
- **Public Opinion Surveys**: Obtaining data from reputable polling organizations or conducting your own surveys can provide structured insights into public opinions on political issues.
- **Election Results:** Historical and real-time election results data can offer valuable information about voting patterns, political affiliations, and changes in political landscapes.

Data Procurement -2-

- **Government Reports and Publications:** Accessing official government reports and publications can provide reliable data on policies, social and economic factors, and government initiatives that influence political orientations.
- **Demographic Data:** External datasets containing demographic information, such as age, gender, income, and education levels, can help correlate political preferences with specific demographic groups.
- **Historical Political Events Data:** Understanding the historical context of political events, protests, and movements through external datasets can contribute to a more nuanced analysis of societal political orientations.
- **International Relations Data:** If relevant to your analysis, data on international relations, treaties, and global events can provide context for understanding how global factors influence local political orientations.
- **Social and Economic Indicators:** External datasets that include social and economic indicators, such as unemployment rates, GDP growth, and income inequality, can offer insights into the broader societal context influencing political orientations.

Data Procurement Challenges

- A considerable **budget** may 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.
- 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.

10

Big Data Privacy Concerns -1-

Imagine a smart city initiative that aims to optimize transportation systems and energy usage through the collection and analysis of telemetry data.

- Data Source 1 (**Car GPS Logs**): The smart city project collects GPS logs from connected cars to analyze traffic patterns, identify congestion points, and optimize transportation routes.
- Data Source 2 (**Smart Meter Data Readings**) Smart meters installed in residential areas provide real-time data on energy consumption, helping the city manage energy resources more efficiently.

Big Data Privacy Concerns -2-

Privacy Implications:

- **Location Tracking:** Over an extended period, the GPS logs from individual cars can create a detailed map of a person's daily movements. This includes home and work locations, places visited, and travel routines.
- **Behavioral Insights:** Analysis of smart meter data can reveal detailed insights into residents' daily routines. Patterns of energy usage indicate when individuals are at home, asleep, or away, providing a snapshot of their lifestyle.
- **Identification Risks:** When combined, GPS logs and smart meter data may lead to the identification of specific individuals. For example, correlating a home address from smart meter data with frequently visited locations from GPS logs can potentially identify the resident.

Big Data Security Challenges -1-

Consider a large e-commerce platform that employs a NoSQL database to manage and analyze vast amounts of customer data.

- E-commerce Platform with NoSQL Database: The NoSQL database is utilized to store and process customer data, product information, and transaction records. This includes details such as customer names, addresses, purchase histories, and payment methods.
- User Access Levels: The platform involves multiple user roles, including customers, support staff, administrators, and data analysts. Each role requires different levels of access to the database based on their responsibilities.

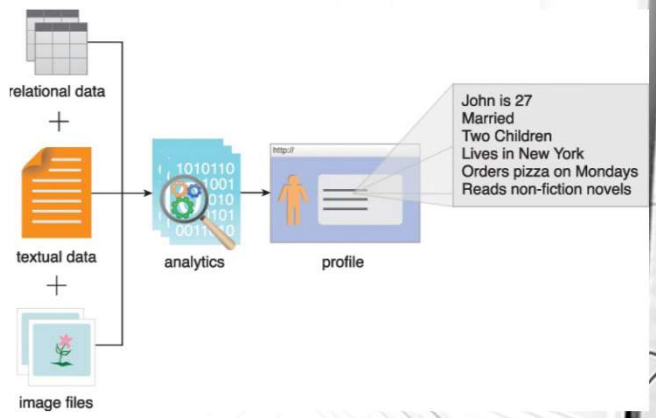
Big Data Security Challenges -2-

Security Challenges:

- Limited Built-in Security Mechanisms: Unlike traditional relational databases with robust security features, many NoSQL databases, especially those using simple HTTP-based APIs, may lack comprehensive built-in security mechanisms. This makes them open to unauthorized access.
- Access Control Issues: Establishing and enforcing granular access controls for different user categories becomes challenging in some NoSQL environments. Inadequate access controls could lead to unauthorized users gaining access to sensitive customer information.

Privacy and Security Considerations

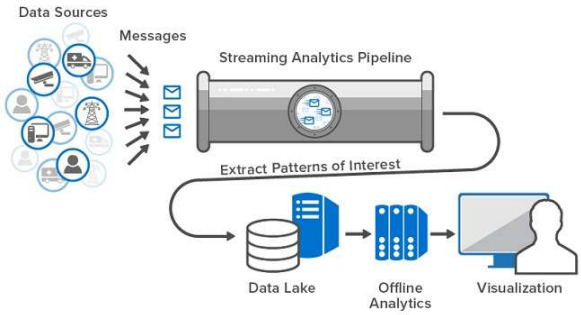
- Performing analytics can reveal confidential information, necessitating privacy safeguards.
- Big Data security involves robust access control and data security measures.



15

Real-time Challenges

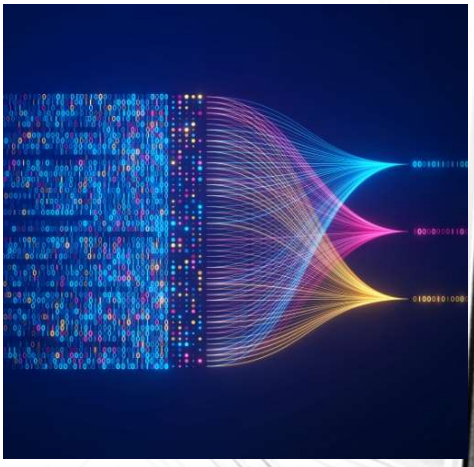
- Some data applications need real-time updates, but many open source Big Data tools work in batches.
- Newer tools support real-time data analysis, but many are proprietary.
- Approaches that achieve near-realtime results often process transactional data as it arrives and combine it with previously summarized batch-processed data.



16

Performance Concerns

- Big Data solutions often face performance concerns due to large data volumes.
- For instance, complex searches on massive datasets can lead to slow query times.
- Another challenge involves network bandwidth, transferring large data amounts may take longer than processing the data itself.
- For example, transferring 1 petabyte of data through a 1-Gigabit LAN connection at 80% throughput can take around 2,750 hours.



17

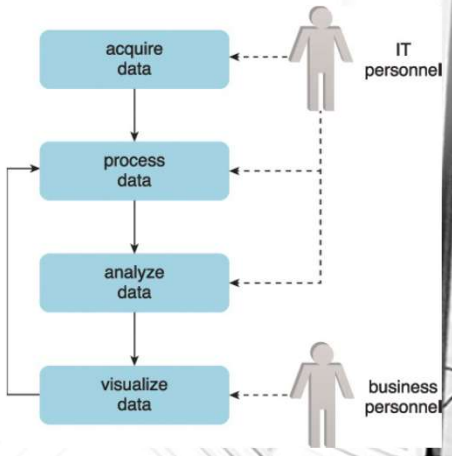
Governance Requirements and Methodology

- Big Data generates valuable assets for a business, so it needs rules.
- A governance framework ensures data and the solution environment follow controlled standards.
- Example of Big Data Governance Framework:
  - Data **Tagging** Standards: The governance framework defines standardized rules for tagging data to ensure consistency and facilitate efficient data management.
  - **Acquisition of External Data:** Governance policies outline the criteria for acquiring external data.
  - Data **Privacy** Management: Stringent rules are established to manage the privacy of customer data.
  - **Archiving** and Retention Policies: The governance framework includes guidelines for archiving historical data and defining retention periods.
  - Data **Cleaning** and Filtering Guidelines: Standardized processes are defined for data cleaning and filtering to enhance data quality.

18

Methodology

- To control how data moves in and out of Big Data solutions, a methodology is essential.
- It considers feedback loops, allowing refined data processing through iterative approaches.
- For instance, business and IT teams can collaborate in cycles, refining the system by adjusting data preparation and analysis steps. Each repetition fine-tunes processing steps, algorithms, and data models, improving accuracy and delivering greater value to the business.



19

Methodology Example

The methodology moves data through a feedback cycle. For example:

- Records flow from EHR systems into the big data platform daily.
- Doctors analyze trends to spot at-risk patients and refine treatment protocols.
- Every month doctors meet IT to discuss results. This identifies ways to enhance models, like adding new vitals to better predict diabetes outcomes.
- IT makes changes and predictions improve over time. This enhances patient care through more effective, data-driven protocols and interventions.
- Overall, governance and methodology work together continuously to ethically and securely improve analytics, insights, and the quality of healthcare over time.

Cloud Adoption for Big Data

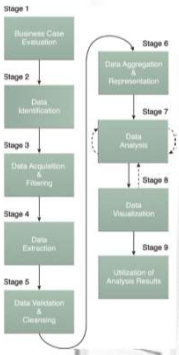
- Cloud adoption may be necessary for scalable environments and cost-effectiveness.
- Resource Constraints: In-house hardware limitations.
- Financial Constraints: Lack of upfront capital for system procurement.
- Isolation Needs: Project isolation to avoid impacting existing business processes.
- Proof of Concept: The Big Data initiative is in the experimental stage.
- Data Residency: Datasets already residing in the cloud.
- Resource Limits: In-house Big Data solution reaching computing and storage limits.

21

Big Data Analytics Lifecycle Overview

Big Data Analytics Lifecycle Overview

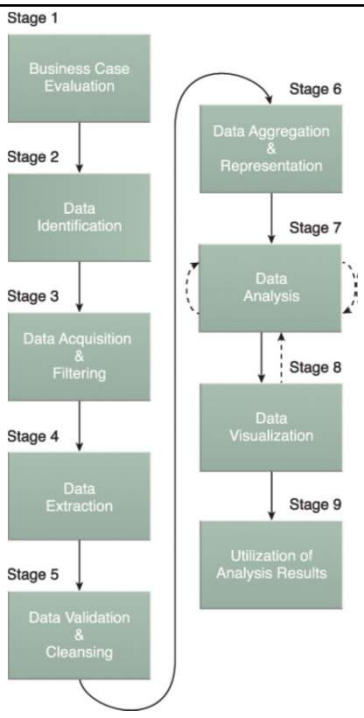
- Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes.
- Introduction of the Big Data analytics lifecycle with nine stages.
  - Include business case evaluation, data identification, extraction, validation, cleaning, aggregation, analysis, visualization, and utilization.
  - Importance of considering training, education, tooling, and staffing in the planning process.



23

The nine stages of the Big Data analytics lifecycle.

Personalized Movie Suggestions on Netflix



Business Case Evaluation

- Crucial starting point for any Big Data analytics lifecycle.
- Involves creating, assessing, and approving a well-defined business case.
- Focus on justification, motivation, and goals of the analysis.
- Identification of Key Performance Indicators (KPIs) for evaluating analytic results.
- Determines if the business problems align with Big Data characteristics (volume, velocity, variety).
- Establishes the budget required for the analysis project.

25

Business Case Evaluation

- Netflix begins by evaluating the business case for personalized movie suggestions.
- The goal is to enhance user experience and increase viewer engagement.
- They analyze the potential benefits of offering smart recommendations based on user preferences.

26

Data Identification

- Dedicated to identifying datasets needed for the analysis and their sources.
- Explores both internal and external data sources.
- Internal datasets may include data marts and operational systems.
- External datasets may involve third-party data providers and publicly available datasets.
- Emphasis on compiling lists, matching against specifications, and identifying a variety of data sources.

27

Data Identification

For Netflix, include:

- viewing history
- day and time of viewing
- device information
- location data
- search queries
- playback interactions
  - (pauses, rewinds, fast-forwards)
- browsing patterns
- completion times for movies or TV shows.

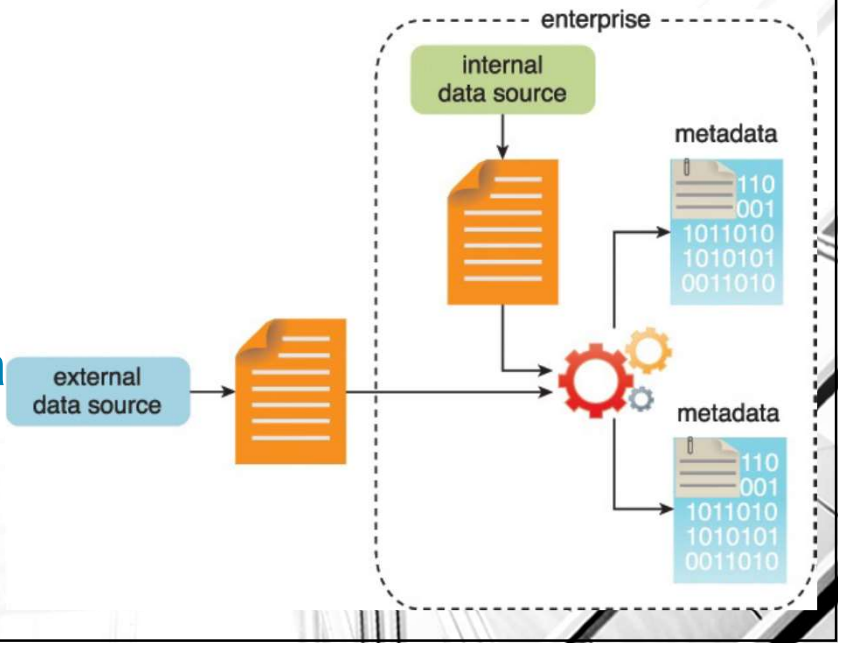
28

Data Acquisition & Filtering

- Focuses on gathering data from identified sources.
- Involves automated filtering to remove corrupt or irrelevant data.
- Data may come in various formats, requiring different acquisition methods.
- Filtering process includes discarding irrelevant data and preserving verbatim copies.
- Metadata addition enhances classification and provenance tracking.

29

Metadata is added to data from internal and external sources



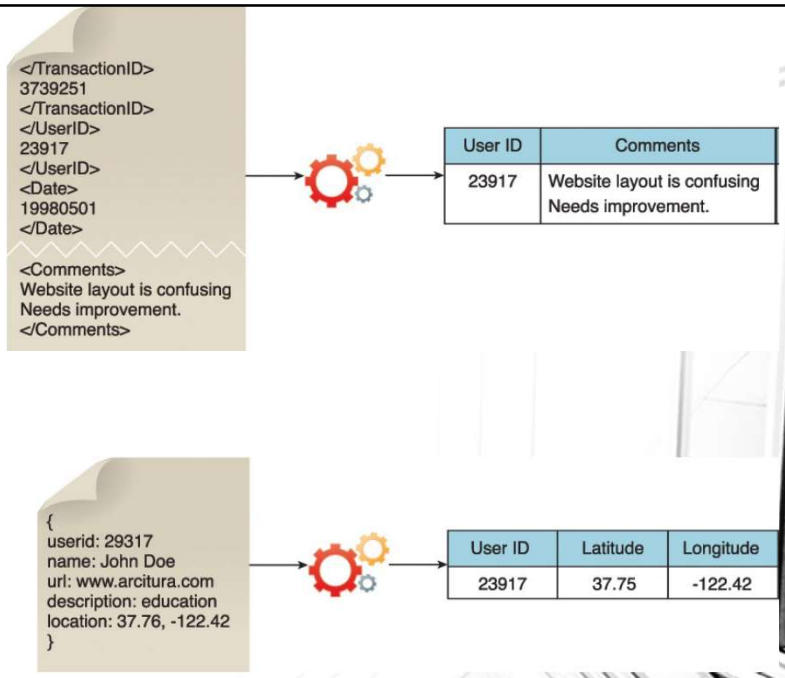
Data Extraction

- Dedicated to extracting and transforming data into a usable format.
- Ensures compatibility with the underlying Big Data solution.
- Extent of extraction depends on the data source and analytic requirements.
- Importance of considering the capabilities of the Big Data solution.
- **Scenario: Improving Patient Outcomes in a Hospital**
  - Structured data tables and text notes migrated from sources

31

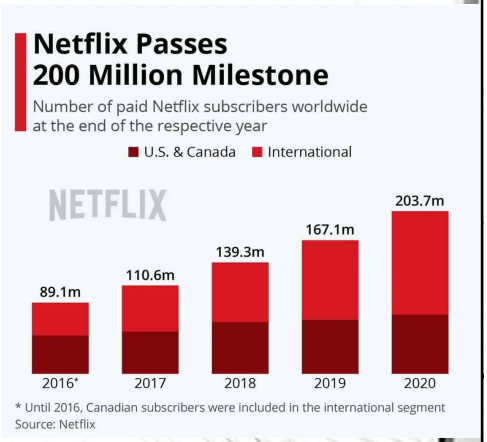
Data Extraction

- Comments and user IDs are extracted from an XML document.
- The user ID and coordinates of a user are extracted from a single JSON field.



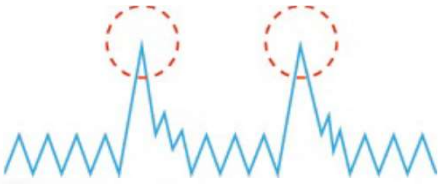
Data Extraction

- Netflix extracts the identified data from its vast pool of over 150 million subscribers.
- This involves gathering information on user behavior, preferences, and interactions with the platform.
- Data is collected in real-time to ensure the most up-to-date insights.



Data Validation & Cleansing

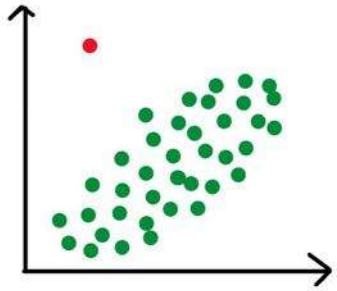
- Establishes validation rules and removes known invalid data.
- Essential to make sure the analysis results are accurate and truthful.
- Differentiates between batch and real-time analytics validation approaches.
- Highlights the role of provenance in determining data accuracy.
- Acknowledges the potential value in seemingly invalid data.



34

Data Validation & Cleansing

- The collected data is validated to ensure accuracy and reliability.
- Netflix employs validation rules to handle discrepancies and outliers.
- Cleaning processes address issues like missing values or inconsistent data, improving the overall quality of the dataset.
- If one dataset lacks certain client information, redundant data from another dataset with similar or matching records may be used to fill in the missing details.



35

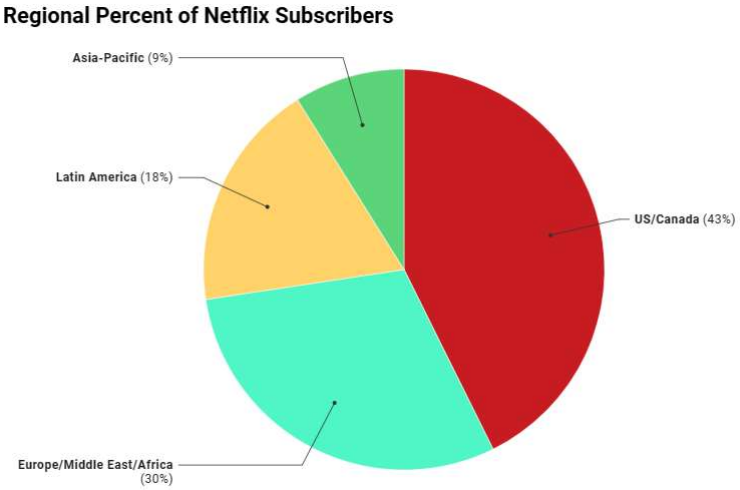
Data Aggregation & Representation

- Integrates multiple datasets for a unified view.
- Addresses challenges related to data structure, semantics, and large volumes.
- Reconciliation of differences requires automated execution.
- Considers future data analysis requirements for fostering reusability.
- Importance of standardized data structures for versatile analysis.

36

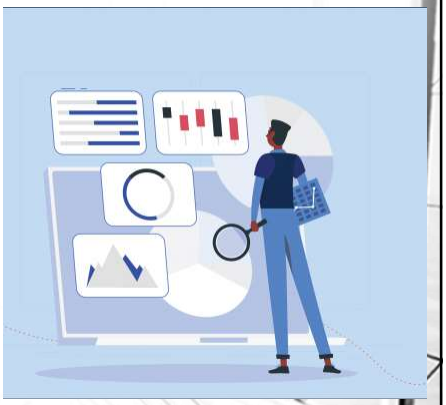
Data Aggregation & Representation

- The extracted and cleaned data is aggregated to form a comprehensive user profile.
- Different datasets, including viewing habits, search history, and interaction patterns, are integrated.
- This stage also involves representing data in a standardized format suitable for analysis.



Data Analysis

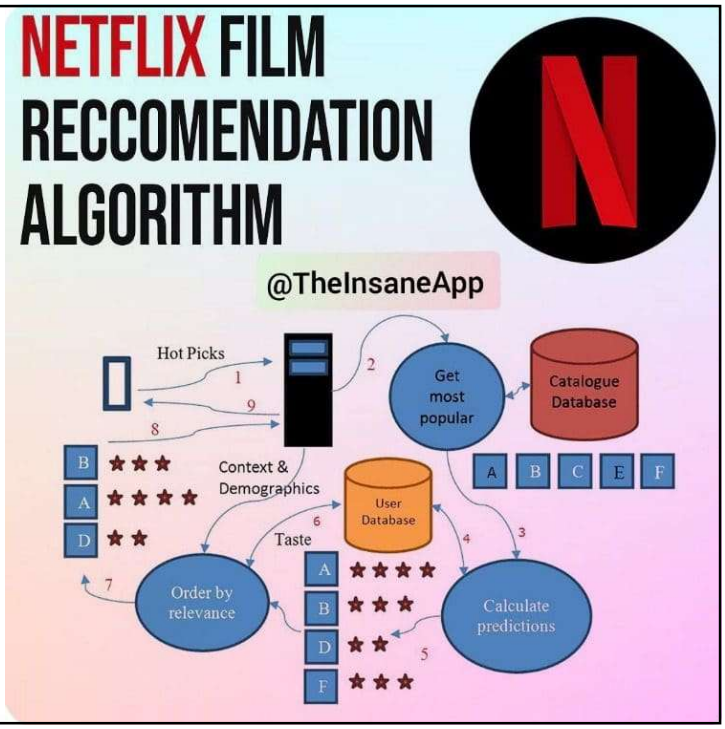
- The Data Analysis stage is all about diving into the collected information, and it can take different forms.
- There are two main types of data analysis:
  - **Confirmatory** analysis starts with a hypothesis that we want to prove or disprove. This approach is more deductive, where we use data to answer specific questions.
  - **Exploratory** analysis is like an adventure. We don't start with a fixed idea; instead, we explore the data to understand what might be going on. it guides us in a general direction and helps us discover new patterns or anomalies along the way.



38

Data Analysis

- Advanced analytics algorithms are applied to the aggregated data to predict user preferences.
- Netflix employs machine learning and recommendation algorithms to analyze viewing habits, search queries, and interaction patterns.
- The analysis aims to forecast what users are likely to watch next.



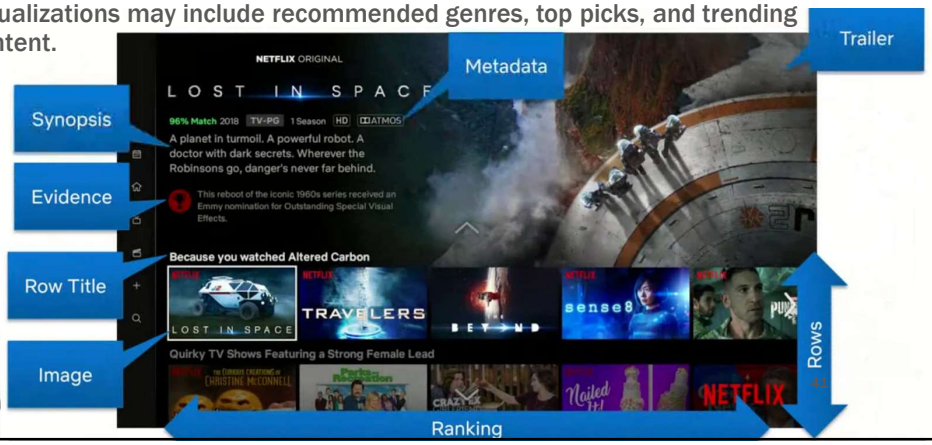
Data Visualization

- Utilizes visualization techniques for effective interpretation by business users.
- Critical for ensuring analysis results are understandable to a broader audience.
- Employs graphical communication to convey complex insights.
- Enables users to perform visual analysis for further exploration.
- Highlights the importance of choosing suitable visualization techniques.

40

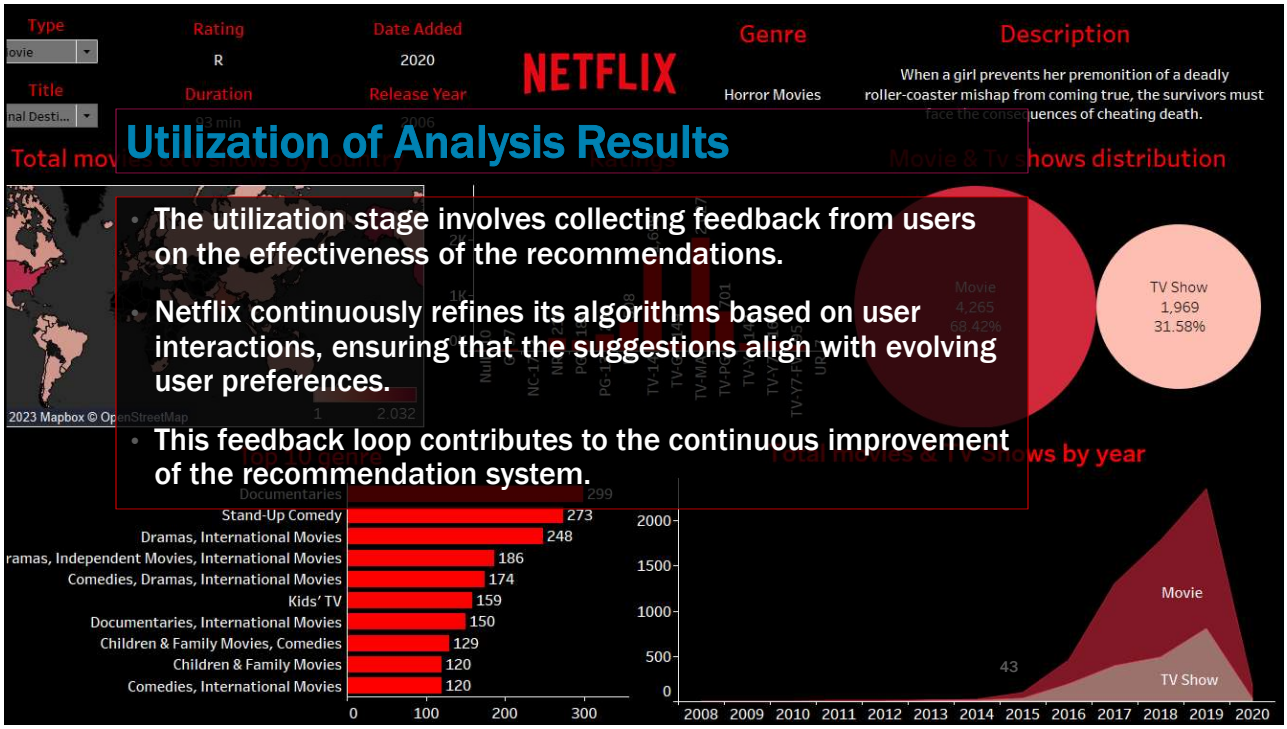
Data Visualization

- Netflix utilizes visualizations to showcase personalized movie suggestions, presenting them in a user-friendly interface within the app.
- Visualizations may include recommended genres, top picks, and trending content.



Utilization of Analysis Results

- This step is about figuring out how and where the analyzed data can be further used.
- In this stage, there are common areas explored:
  - Input for Enterprise Systems: The analysis results can be fed into enterprise systems, automatically or manually, to enhance their behaviors and performance.
  - Business Process Optimization: Patterns and anomalies found during analysis refine business processes. An example is optimizing transportation routes in a supply chain.
  - Alerts: Analysis results can be used to create alerts, either for existing ones or new ones. For instance, users may be informed via email or SMS about an event that requires corrective action based on the analysis results.



Conclusion

- This chapter underscores the transformative potential of Big Data adoption, move organizations to not only navigate the technical aspects but also invest in the people and processes that drive success.
- As we start on this Big Data journey, the lifecycle provides a structured path, emphasizing the importance of data-driven decision-making and the iterative nature of discovering valuable insights in the vast sea of data.