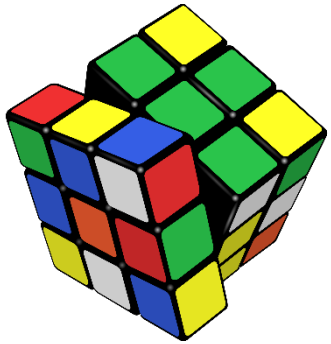


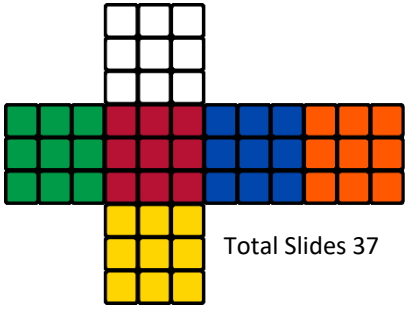
Big Data Engineering For Analytics

Managing the Big Data

Dr. Liu Fan

(isslf@nus.edu.sg)

NUS-ISS



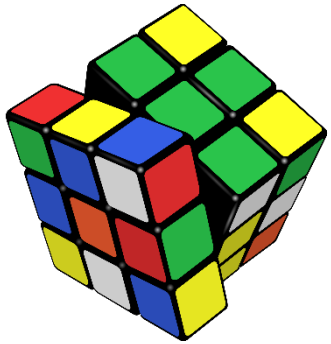
Total Slides 37

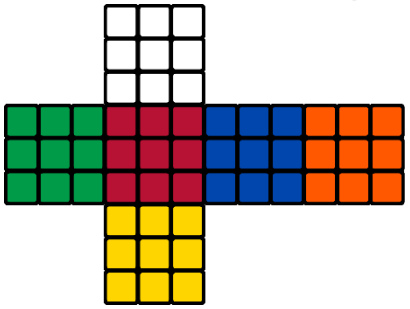
Learning Objectives

- Understand the fundamental concepts and practices involved in managing an analytics project
- Understand the knowledge domains and best practices for each stage of this lifecycle
- Evaluate the various management techniques on context, capability, strengths and shortcomings.

Agenda

- Big Data Management Practices
- Analytics Lifecycle Toolkit
- Summary





Big Data Management Practices

Steven Hawking on the Universe:

"It would not be much of a universe if it wasn't home to the people you love."

Big Data Management

- Big data management incorporates the policies, techniques, and processes used for data collection, storage, administration, and the delivery of large repositories.
- Major management concerns are around:
  1. Data ingestion
  2. Data storage
  3. Data quality
  4. Data operations
  5. Data scalability and security

Big data management services

- Different vendors support different technological stacks, and have different pricing models.
- There are vendors that offer a variety of standalone or multi-featured big data management tools.
- The management can also involve additional cleaning, integration, migration, and reporting.

Data Cleansing

- Data cleansing is the process of identifying and fixing corrupt or fallacious records in a record set, table, or database.
- It also deals with identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data, and then replacing, modifying, or deleting the infected data.
- Data cleansing is important because of:
  - **Heterogeneity** The main data is usually spread across different legacy systems, including spreadsheets, text files, and web pages
  - **Accuracy** By ensuring that the data is as accurate as possible, an organization can maintain good relationships with its customers, improving the organization's efficiency
  - **Completeness** Correct and complete data provides better insights into the process that the data concerns

Data Integration

- Data integration is one of the techniques of combining data from disparate sources and providing end users with a unified view of that data.
- This gives a sense of abstraction to the end users.

Other management capabilities

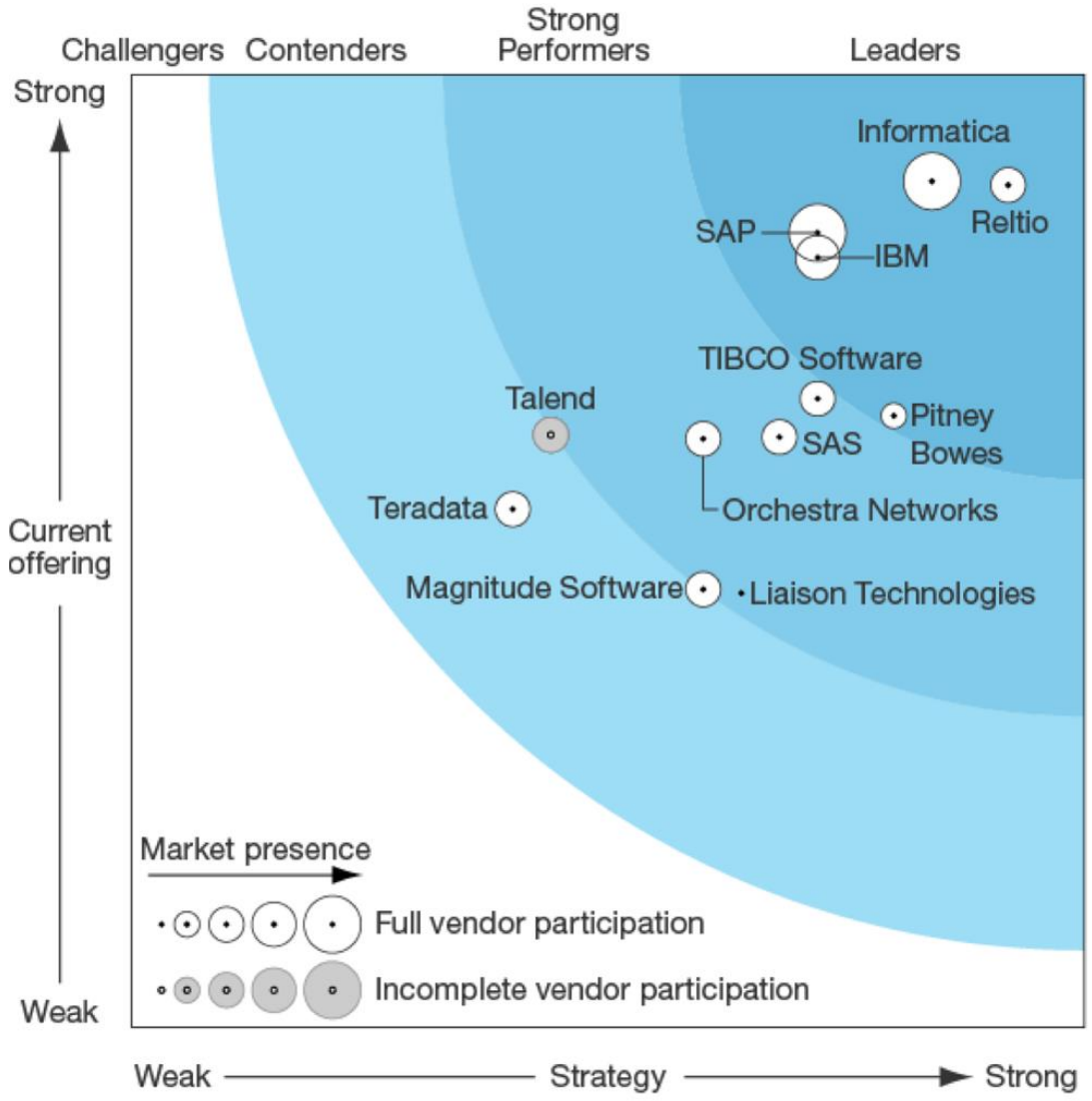
- **Data migration**: This is the process of transferring data from one environment to another. Most migration occurs between computers and storage devices (for example, transferring data from in-house data centers to the cloud).
- **Data preparation**: Data that is used for analysis is often messy and inconsistent, and not standardized. This data must be collected and cleaned into one file or data table, before an actual analysis can take place. This step is referred to as data preparation. It involves handling messy data, trying to combine data from multiple sources, and reporting on the data sources.
- **Data enrichment**: This step involves enhancing the existing set of data by refining the data, in order to improve its quality. It can be done in several ways. Some common ways are by adding new datasets, correcting miniature errors, or extrapolating new information from raw data.
- **Data quality**: This is the act of confirming that the data is accurate and reliable. There are several ways in which data quality is controlled.

Other management capabilities

- **Data analytics**: This is the process of drawing insights from datasets by analyzing them with a variety of algorithms. Most steps are automated by using various tools.
- **Master data management (MDM)**: This is a method that is used to define and manage the important data of any enterprise, in order to facilitate the process of linking critical enterprise data to one master set. The master set works as a single source of truth for the organization.
- **Data governance**: This is a data management concept that deals with the ability of a company to ensure high data quality throughout the analytical process. This process includes warranting the availability, usability, integrity, and accuracy of data.
- **Extract transform load** (**ETL**): As the name implies, this is the process of moving data from an existing repository to a different database, or a new data warehouse.

Vendors

- Alation
- AtScale
- Cloudera & Hortonworks
- Collibra
- Confluent
- SAS
- Verato
- TIBCO
- Talend



Reference: https://www.reltio.com/solutions/master-data-management/

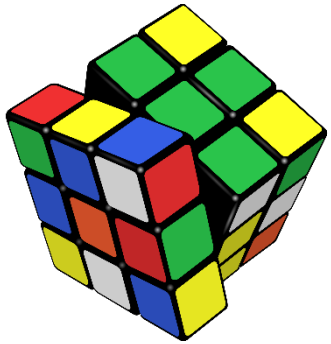
Apache Atlas and Apache Pulsar

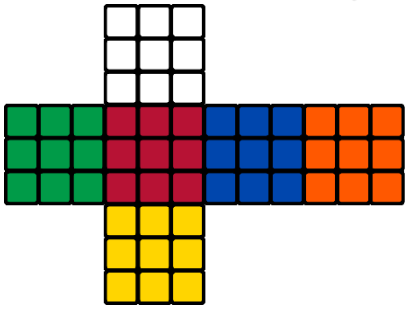
Atlas is a scalable and extensible set of core foundational governance services – enabling enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the whole enterprise data ecosystem.

Apache Atlas

Apache Pulsar is an open-source distributed pub-sub messaging system originally created at Yahoo and now part of the Apache Software Foundation

PULSAR



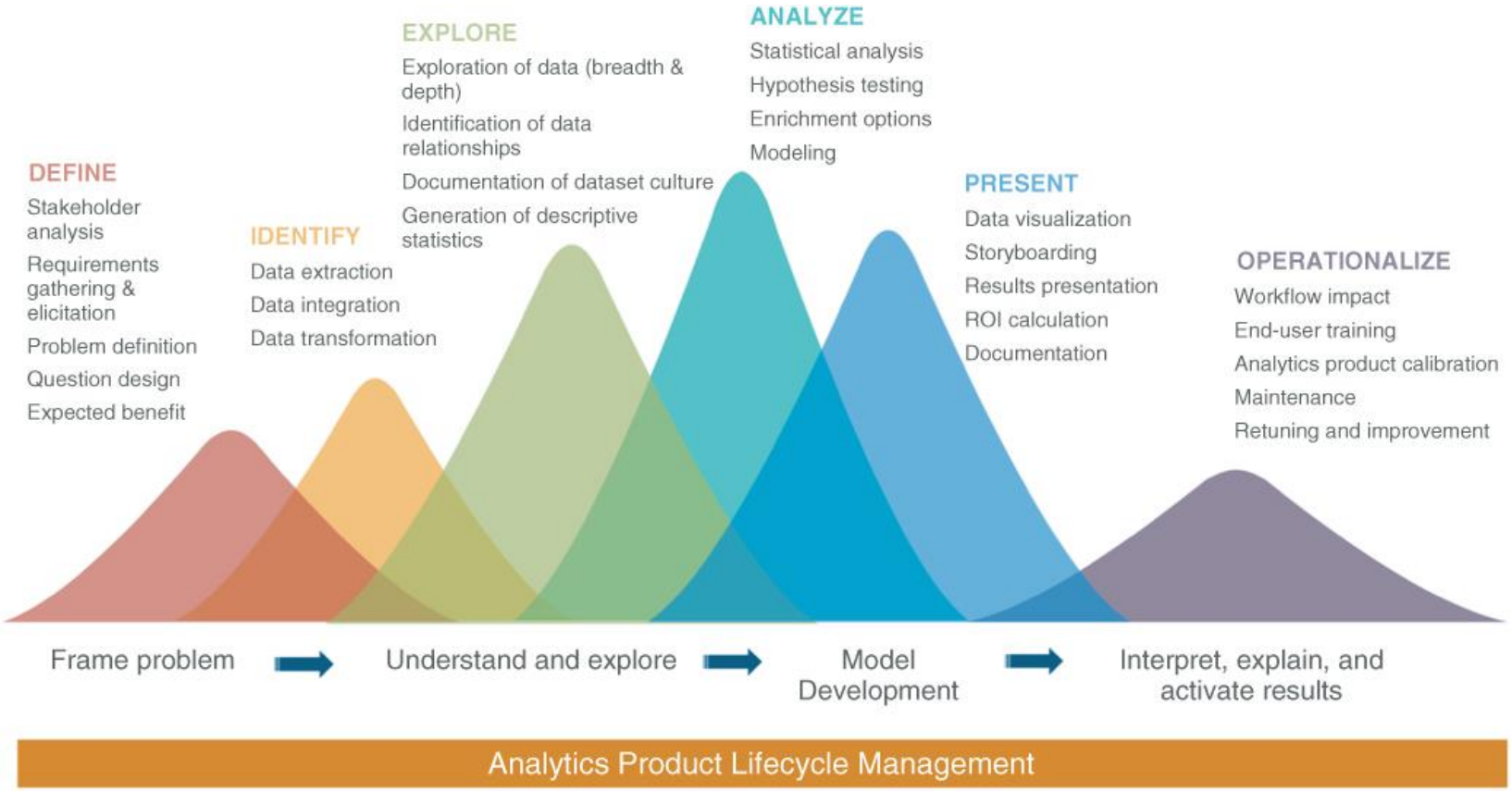


Analytics Lifecycle Toolkit

You never lose a dream. It just incubates as a hobby.

~Larry Page

The Analytics Lifecycle begins with a definition of the fundamental question or problem and continues through data exploration, analysis, and results activation. The loop is closed when analytics insights are operationalized into the business workflow in some way.

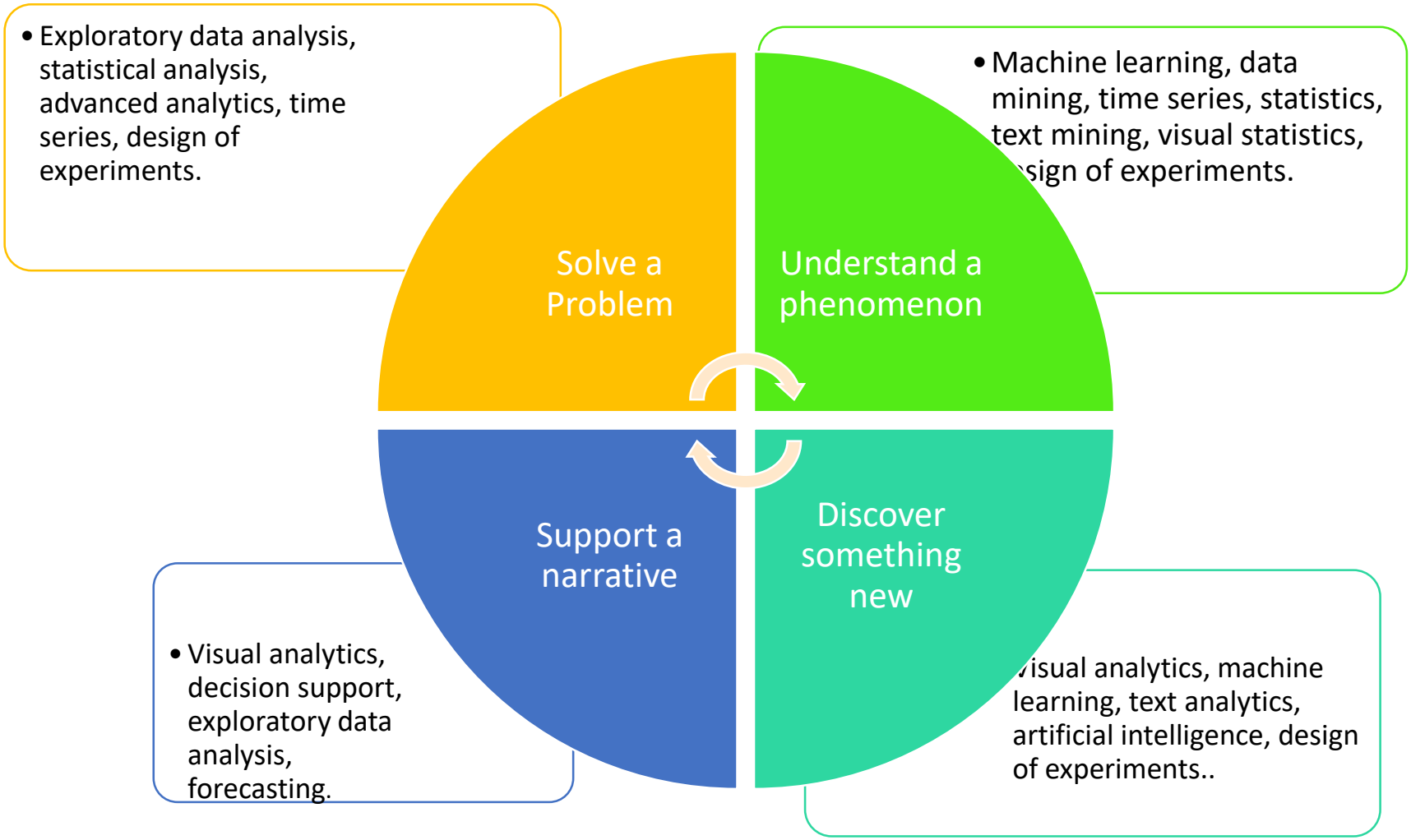


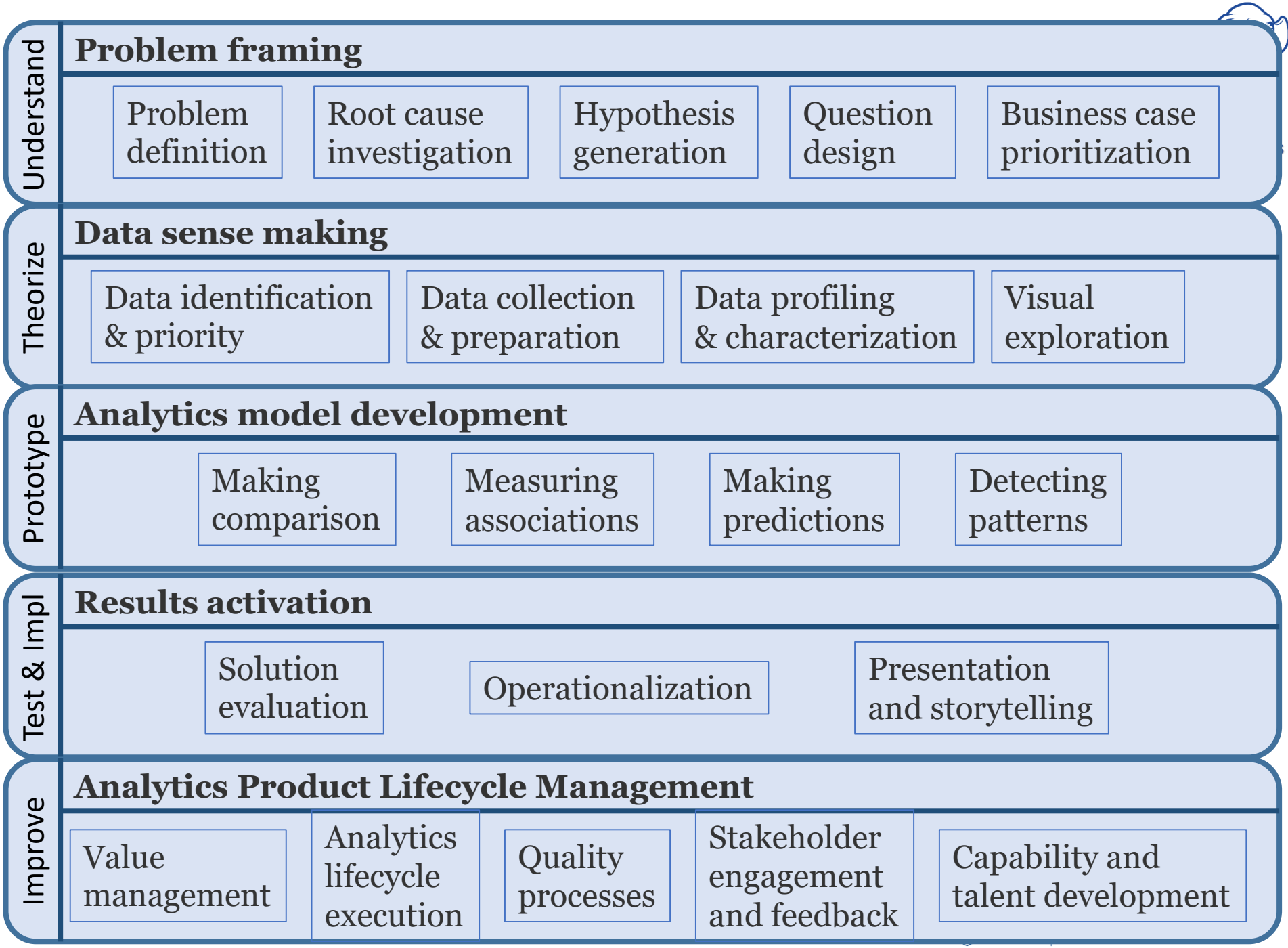
https://support.sas.com/resources/papers/proceedings17/0832-2017.pdf

Goals of Analytics

- **To solve a problem**—Applied analytics is concerned with the discovery of solutions to practical problems, and we measure success by its immediate utility or application. In the Analytics Lifecycle, the emphasis is usually on speed and explanatory value.
- **To support a narrative**—Analytics professionals often use techniques to directly support a story, such as confirming a hypothesis or visualizing a relationship. The outcome is of primary benefit to emphasize accuracy and reliability, and not necessarily the repeatability of the process.
- **To understand a phenomenon**—We often embark on an analytics project to understand a phenomenon more fully in the most general and parsimonious way possible. Techniques such as visual analytics or exploratory data analysis can support the discovery of these relationships.
- **To discover something new**—Analytics can be used to tell us something that we didn't already know. We can trace this objective back to its heritage in data mining; it often takes advantage of new analytics methods or unstructured and unorganized data, and it is often focused on innovation-type endpoints. Curiosity and inquisitiveness motivate the discovery about the relationships and associations.

Motivation for Analytics




Understand
Problem framing
Problem definition
Root cause investigation
Hypothesis generation
Question design
Business case prioritization
Theorize
Data sense making
Data identification & priority
Data collection & preparation
Data profiling & characterization
Visual exploration
Prototype
Analytics model development
Making comparison
Measuring associations
Making predictions
Detecting patterns
Test & Impl
Results activation
Solution evaluation
Operationalization
Presentation and storytelling
Improve
Analytics Product Lifecycle Management
Value management
Analytics lifecycle execution
Quality processes
Stakeholder engagement and feedback
Capability and talent development

Problem framing

Problem definition	• Identify and characterize the business problem/need. • Manage the problem definition and impact. • Support the justification for effort. • Reformulate problem statement as an analytics problem and/or technical requirements. • Identify assumptions related to the problem and proposed solution. • Refine the business and analytics problem statements.
Root cause investigation	• Utilize brainstorming techniques and effectively use divergent thinking processes to uncover potential cause-and-effect relationships. • Classify requirements appropriately and determine feasibility. • Apply root cause analysis to requirement definitions.
Hypothesis generation	• Generate (and manage) testable hypotheses. • Validate expected results and key requirements information with stakeholders. • Generate testable theories and validate their reasonableness. • Shadow workflows that are not understood. • Conduct primary and secondary research as needed to understand potential sources of the issue.

Problem framing

Problem definition	• Utilize the FINER criteria (Car, 2013) to evaluate whether a problem can be translated into a question that can be answered. • Convert a question into a proper study design.
Business case prioritization	• Prioritize requirements based on business value, cost to deliver, and time constraints. • Validate that solution design meets the business need. • Define the capabilities needed to support solution. • Manage the metrics related to solution implementation and success.

Data sense making

Data identification and prioritization	• Articulate the data required to solve the problem. • Reconcile the difference between the data we can get versus data that we want. • Trace back the business and operational workflows reflected in the data. • Articulate the provenance and governance assumptions of the data.
Data collection and preparation	• Extract data from large, structured data stores. • Extract data from unstructured data sources. • Integrate data from multiple sources. • Ensure privacy and protection of data. • Utilize a variety of methods to cleanse and/or enrich data. • Map results back to business and operational workflows. • Model the data appropriately for the type of analysis needed.

Data sense making

Data profiling and characterization	• Identify relationships in the data. • Perform exploration of unknown data. • Profile datasets. • Develop and execute a structured process to describe the aggregate trends, features, and culture of a data set. • Generate descriptive statistics, frequency analysis, and distributions of data (aka, exploratory data analysis – EDA). • Identify and investigate outlier data. • Develop theories that might address the problem.
Visual exploration	• Utilize a variety of programmatic and menu-driven visualization tools to examine associations. • Utilize principles of good design to craft visuals appropriate to their type. • Create graphics that help express the context and insight of the data.

Analytics model development

Making comparisons	• Determine appropriate statistical tests and utilize them in basing conclusions. • Apply a wide variety of statistical models, processes, routines, and measures to compare two or more groups. • Compare and contrast features of categorical and numerical data sets using appropriate tests. • Apply quantitative measures to describe the properties of a sample of data. • Define and apply statistical significance, confidence intervals, effect size, and hypothesis testing. • Differentiate between categorical versus continuous data and the appropriateness of various testing strategies used for making inferences.

Analytics model development

Measuring associations	• Utilize visualization methods to examine relationships between different types of data. • Distinguish between an explanatory and response variable and their role in tests of association. • Describe the types of tests used in measuring associations including those in parametric and non-parametric testing. • Relay the difference between an association and a cause-and-effect relationship.
Making predictions	• Identify the two classes of prediction models. • Enumerate the types and methods of supervised and unsupervised methods used for prediction models. • Relate the type of prediction problem being asked back to the methods available in statistics, data mining, and machine learning. • Recognize common analytics methods such as predictive models, cluster analysis, neural networks, and machine learning.
Detecting patterns	• Classify the types of problems that we can solve using pattern recognition. • Describe the various classification approaches. • Illustrate the difference between feature selection and feature extraction. • Describe the difference between classification and discrimination.

Results activation

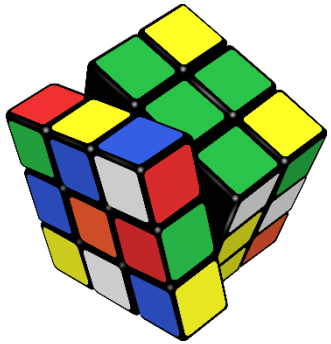
Value management	• Conduct data/analytics output interpretation. • Coach and mentor stakeholders. • Perform business validation of the model. • Compare results from various models. • Explore alternative explanations.
Operationalization	• Incorporate a set of analytics and insights into business workflow such that a continual, positive benefit is seen and the organizational learning paradigm is realized. • Create model, usability, and system requirements for production. • Deliver production model. • Support the business process change. • Support the implementation of the model. • Assess actionability and impact to operational workflows. • Document and communicate findings (including assumptions, limitations, and constraints).
Presentations and storytelling	• Communicate effectively with various audiences. • Create data visualizations that convey meaning. • Deliver report with findings. • Evangelize value of analytics/business benefits. • Socialize analytics results, advances.

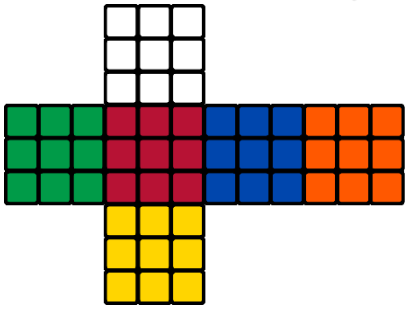
Analytics Product Lifecycle management

Value management	• Strategic alignment. • Develop and support a collaborative product management culture. • Analytics evangelism. • Evaluate the business benefit of analytics over time.
Analytics lifecycle execution	• Analytics prioritization. • Utilize project management principles to define, execute, and manage project activities. • Develop and establish project goals and milestones. • Implement project standards and procedures. • Monitor and analyze project costs. • Estimate project work. • Manage capital and expense reporting.
Quality processes	• Improve the way in which data is governed. • Create and follow quality management plans. • Promote continuous improvement. • Follow robust testing and quality processes. • Utilize a risk-based validation approach. • Participate in peer reviews of both code and data products. • Document and improve quality principles. • Track model quality and durability. • Recalibrate and maintain models.

Analytics Product Lifecycle management

Stakeholder engagement and feedback	• Analyze impact of change. • Support training and communication activities. • Manage change. • Promote processes for managing change and measuring the impact of decisions • Promote a culture of sharing and collaboration.
Capability and talent development	• Manage resources and evaluate performance. • Manage portfolio of projects with available resources. • Manage talent development. • Provide performance improvement coaching. • Manage conflict. • Demonstrate leadership and influence with the team and with external stakeholders. • Assess lessons learned. • Catalog data assets and ensure that metadata is accessible and usable.





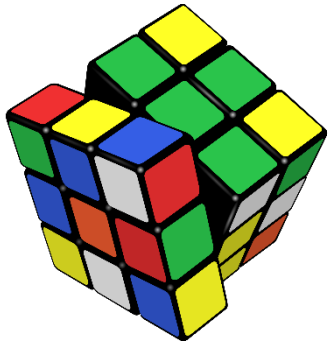
Summary

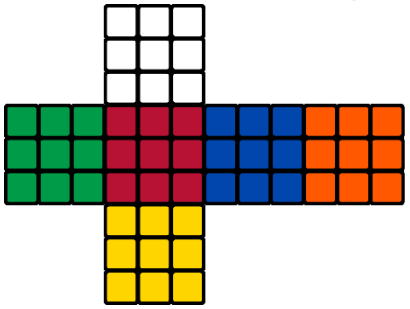
"What is the one sentence summary of how you change the world? Always Work hard on something uncomfortably exciting!"

~Larry Page

Essential Points

- Big Data analysis blends traditional statistical data analysis approaches with computational ones.
  - The overall goal of data analysis is to support better decision-making.
  - Carrying out data analysis helps establish patterns and relationships among the data being analysed.
- The Big Data analytics lifecycle can be divided into the following nine stages.
  - Business Case Evaluation; Big Data Identification; Big Data Acquisition & Filtering; Big Data Extraction; Big Data Validation & Cleansing; Big Data Aggregation & Representation; Big Data Analysis; Big Data Visualization; Big Data Utilization of Analysis Results;
- There are many analysis techniques such as Quantitative analysis, Qualitative analysis, Data mining, Statistical analysis, Machine learning, Semantic analysis and Visual analysis





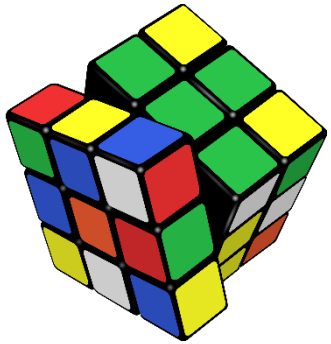
References

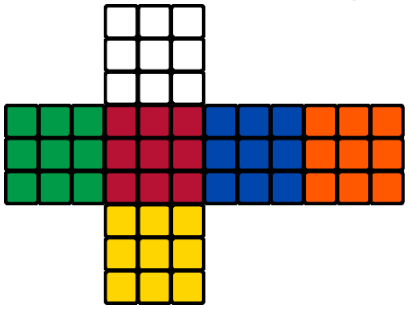
"If you're not doing some things that are crazy, then you're doing the wrong things."

~Larry Page

References

- The Analytics Lifecycle Toolkit, by Gregory S. Nelson, Published by Wiley, 2018
- Big Data Fundamentals: Concepts, Drivers & Techniques by Thomas Erl; Wajid Khattak; Paul Buhler, Published by Prentice Hall, 2016
- Hadoop Essentials by Swizec Teller, Published by Packt Publishing, 2015.
- Data Science from Scratch by Joel Grus, Published by O'Reilly Media, Inc., 2015
- Designing Data-Intensive Applications, 1st Edition, by Martin Kleppmann, Published by O'Reilly Media, Inc., 2017
- Data Analytics with Hadoop, by Jenny Kim; Benjamin Bengfort, Published by O'Reilly Media, Inc., 2016
- YARN Essentials by Amol Fasale; Nirmal KumarPublished by Packt Publishing, 2015
- Big Data Bootcamp: What Managers Need to Know to Profit from the Big Data Revolution by David Feinleib*Published by Apress, 2014*
- Instant Apache Sqoop by Ankit Jain *Published by Packt Publishing, 2013*



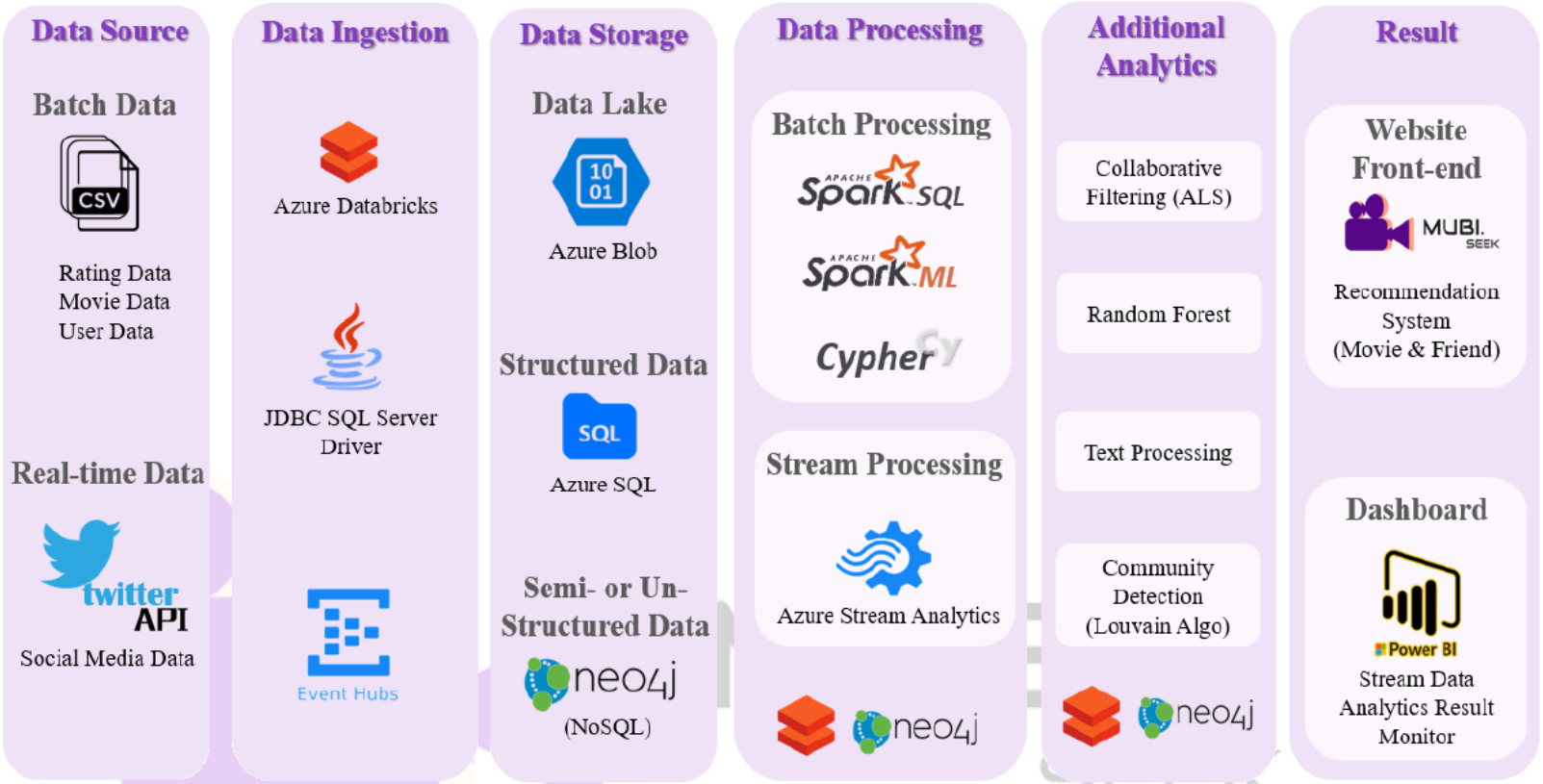


Appendix

MUBI Movie Recommendation System

- Build a recommendation system which can recommend MUBI users with their most-interested-and-loved movies, providing personalized experience.
- Combine with social media analysis (Twitter streaming data) to track people's current points of interest and popularity of movies, trying to reinforce the movie recommendation system.
- Combine with text analytics to predict the lack rating score from a user on a movie based on the user's review text, trying to reinforce the movie recommendation system.

Overall Architecture



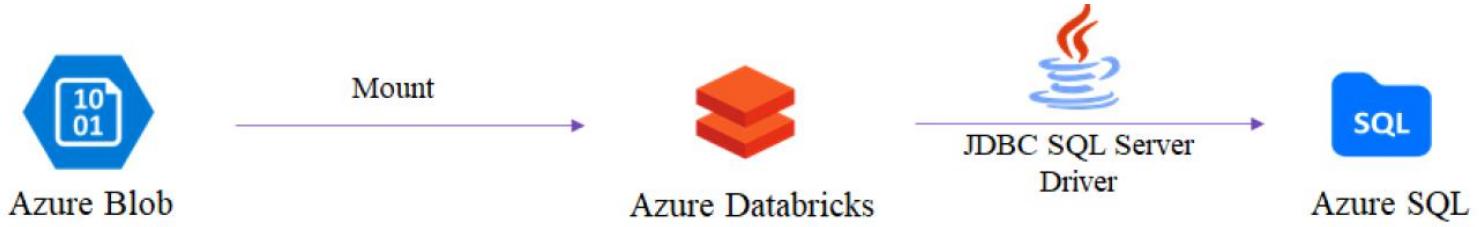
Data Sources

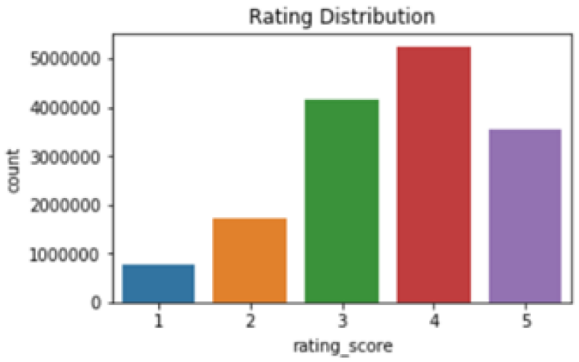
- The first dataset is from Kaggle website, called MUBI SVOD Platform Database for Movie Lovers

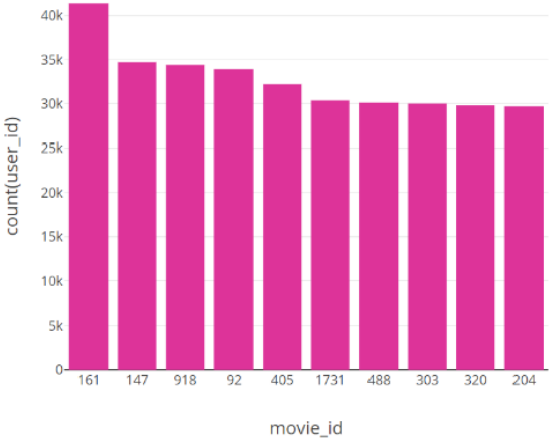
- The second data source is real-time social media data (text messages) from Twitter API (https://developer.twitter.com/en/docs/twitter-api), which will be used for social media analysis to track current points of interest and popularity of movies.

Batch data ingestion and processing

Big Data Engineering For Analytics



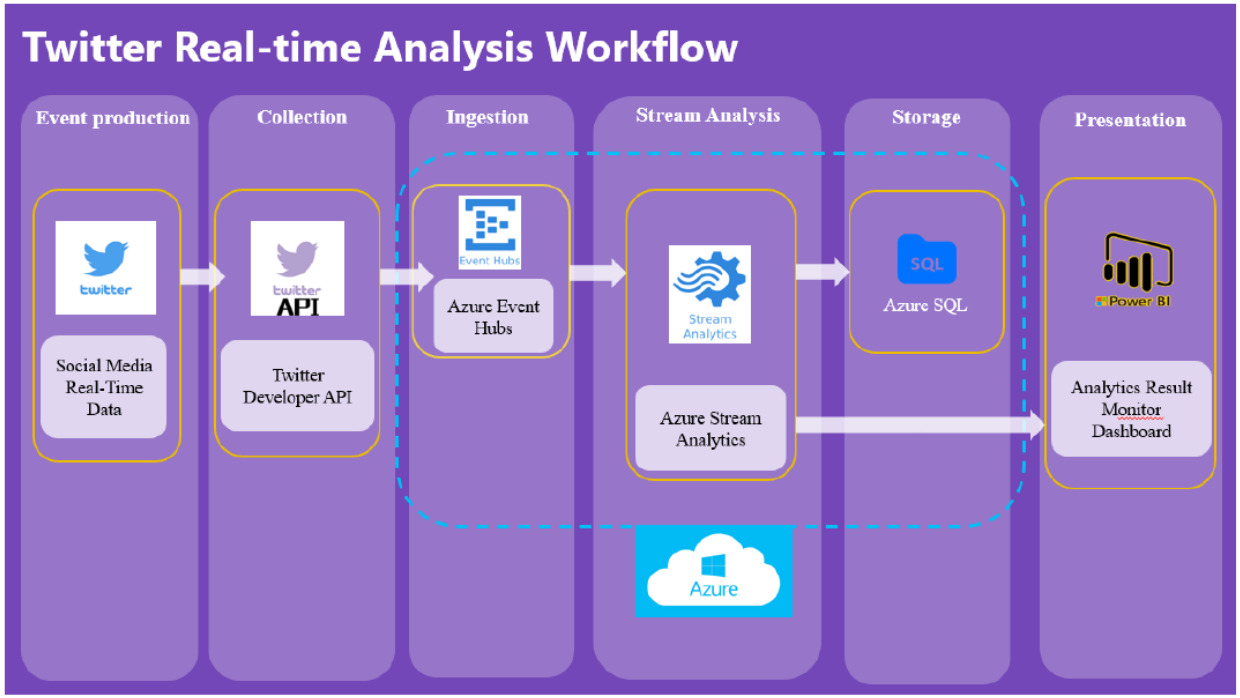




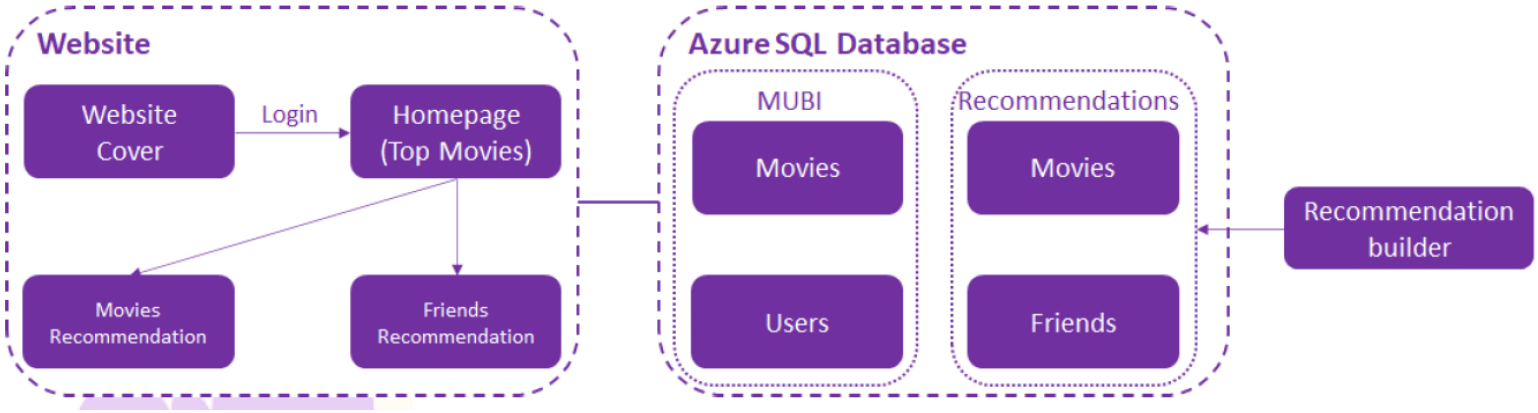
Twitter streaming data processing

Movie Recommendation System

- Collaborative Filtering
- Text processing (Reviews)
- Social Media Analysis(Twitter Streaming Data)



Web Front-end Deployment



Used Django framework to build a demo website to show the results of movies and movie recommendation.