05 Big Data Life Cycle - KirkYagami

Big Data Analytics Life Cycle: In-Depth Lecture Notes

I. Introduction

The Big Data Analytics Life Cycle is a comprehensive process for handling and analyzing large, complex datasets. It consists of nine phases, each crucial for turning raw data into actionable insights.

II. The Nine Phases

1. Business Case/Problem Definition

- Objective: Understand the business context and define the problem.
- Key Activities:
  - Learn about the business domain
  - Identify specific problems or opportunities
  - Frame the business problem as an analytics challenge
  - Estimate potential gains and required resources
  - Determine if it's truly a big data problem (volume, velocity, variety)
- Importance: Sets the direction for the entire project

2. Data Identification

- Objective: Locate appropriate datasets for analysis.
- Key Activities:
  - Research similar cases in other companies
  - Identify internal data sources (e.g., feedback forms, existing software)
  - Explore external data sources (e.g., third-party providers)
- Consideration: Data relevance to the business case is crucial

3. Data Acquisition and Filtration

- Objective: Gather and initially clean the data.
- Key Activities:
  - Collect data from identified sources
  - Remove corrupt or irrelevant data
  - Store a compressed copy of filtered data for potential future use
- Challenge: Dealing with mostly unstructured data

4. Data Extraction

- Objective: Ensure all data is compatible with the analysis scope.
- Key Activities:
  - Identify incompatible data entries
  - Transform incompatible data to fit the analysis requirements
- Importance: Prepares data for more detailed cleaning and validation

5. Data Munging (Validation and Cleaning)

- Objective: Thoroughly clean and validate the data.
- Key Activities:
  - Remove invalid data
  - Establish and apply complex validation rules
  - Handle null entries (e.g., fill from similar datasets or remove)
- Importance: Ensures data quality for accurate analysis

6. Data Aggregation & Representation (Storage)

- Objective: Combine datasets and prepare for storage.
- Key Activities:
  - Join multiple datasets using common fields
  - Consider automation for large-scale operations
- Challenge: Handling potentially very large amounts of data

7. Exploratory Data Analysis

- **Objective**: Analyze the data to extract insights.
- **Types of Analysis**:
  1. Confirmatory Analysis:
     - Test pre-existing hypotheses
     - Provide definitive answers to specific questions
  2. Exploratory Analysis:
     - Discover patterns and relationships in the data
     - Answer "why" a phenomenon occurred
- **Importance**: Core step where insights are generated

8. Data Visualization (Preparation for Modeling and Assessment)

- **Objective**: Represent findings visually for easy interpretation.
- **Key Activities**:
  - Use visualization tools to create graphics
  - Ensure visualizations are understandable to business users
- **Benefits**:
  - Aids in result interpretation
  - Can reveal answers to unasked questions

9. Utilization of Analysis Results

- **Objective**: Apply insights to business decisions.
- **Key Activities**:
  - Make data-driven decisions
  - Optimize and refine business processes
  - Use results as input for enhancing system performance
- **Importance**: Translates analysis into tangible business value

III. Iterative Nature of the Process

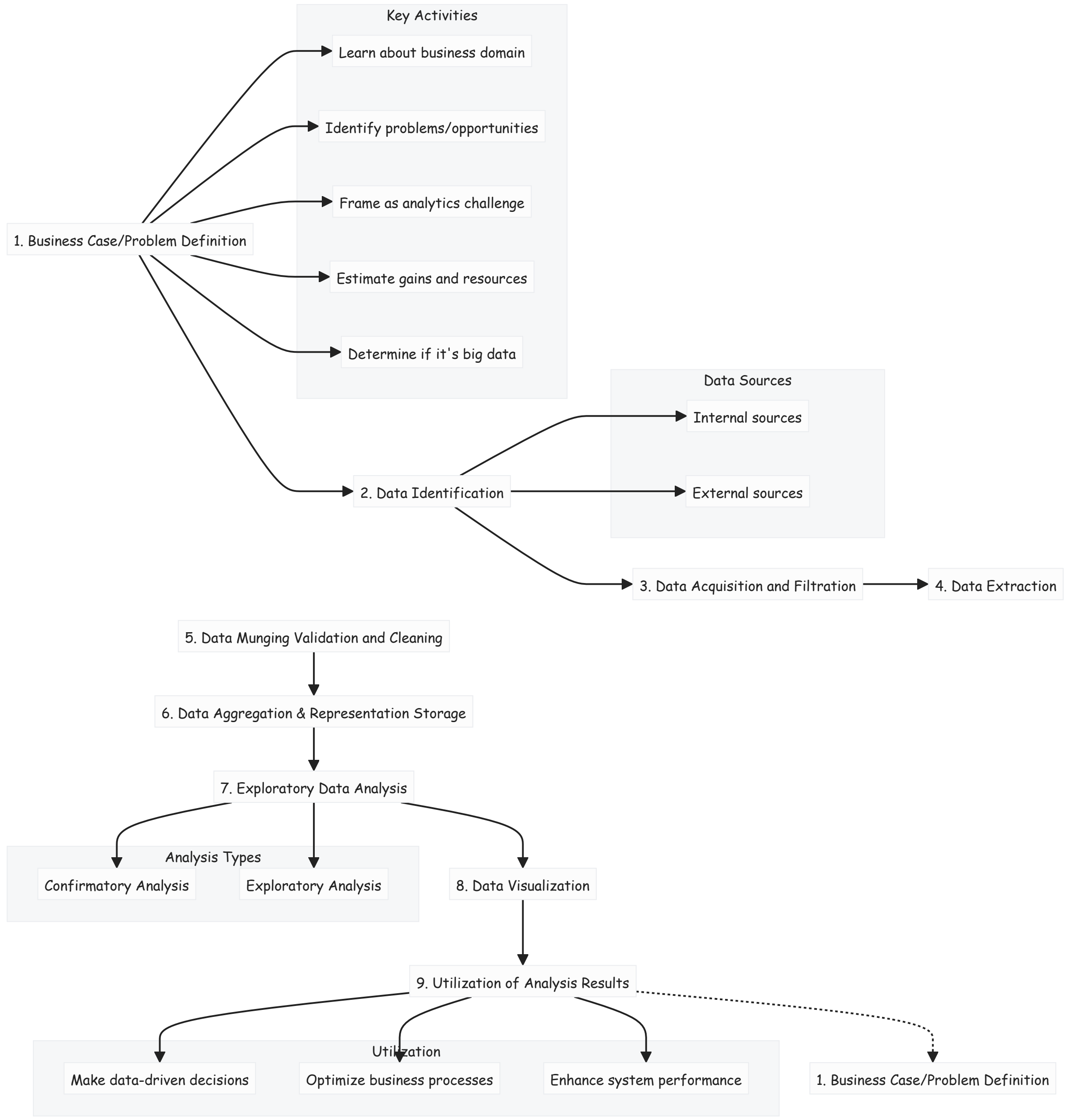
- Phases 7 and 8 (Analysis and Visualization) may be repeated for better results
- Emphasis on error correction and refinement
- Allows for moving back from Phase 8 to Phase 7 if needed

IV. Key Considerations

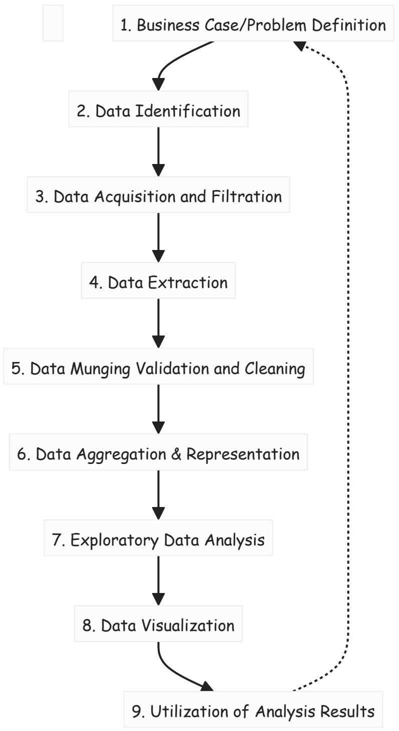
1. **Data Quality**: Crucial throughout the process, especially in phases 3-5
2. **Scalability**: Must handle large volumes of data efficiently
3. **Expertise**: Requires a mix of business knowledge and technical skills
4. **Technology**: Appropriate tools needed for each phase, especially for analysis and visualization
5. **Privacy and Ethics**: Important when dealing with sensitive data

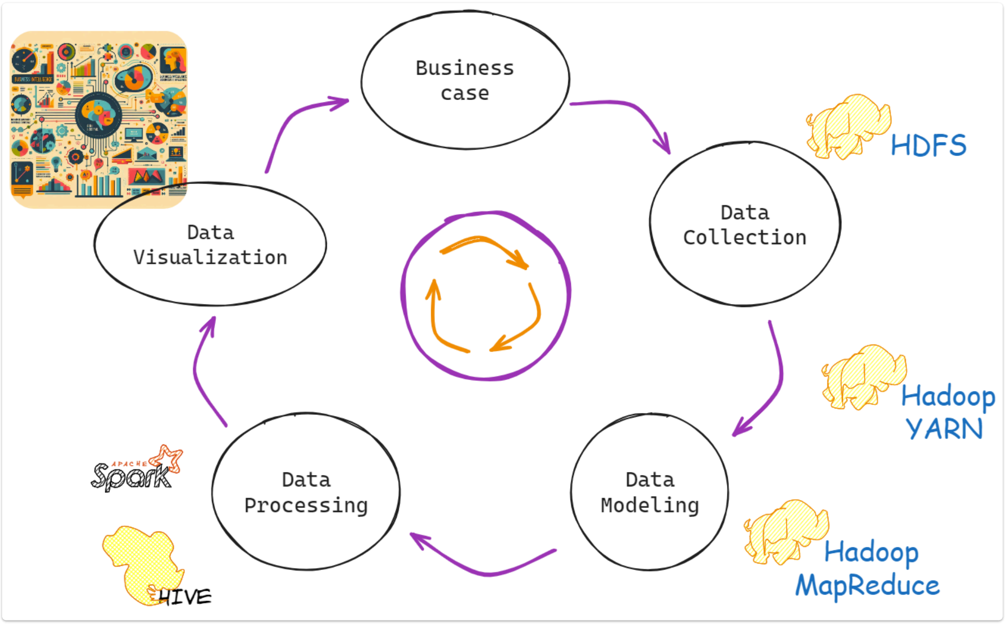
V. Conclusion

The Big Data Analytics Life Cycle is a comprehensive approach to extracting value from complex datasets. It requires careful planning, rigorous data handling, and skilled analysis to turn raw data into actionable business insights.


Key Activities
Learn about business domain
Identify problems/opportunities
Frame as analytics challenge
1. Business Case/Problem Definition
Estimate gains and resources
Determine if it's big data
Data Sources
Internal sources
2. Data Identification
External sources
3. Data Acquisition and Filtration
4. Data Extraction
5. Data Munging Validation and Cleaning
6. Data Aggregation & Representation Storage
7. Exploratory Data Analysis
Analysis Types
Confirmatory Analysis
Exploratory Analysis
8. Data Visualization
9. Utilization of Analysis Results
Utilization
Make data-driven decisions
Optimize business processes
Enhance system performance
1. Business Case/Problem Definition

Data Life Cycle


1. Business Case/Problem Definition
2. Data Identification
3. Data Acquisition and Filtration
4. Data Extraction
5. Data Munging Validation and Cleaning
6. Data Aggregation & Representation
7. Exploratory Data Analysis
8. Data Visualization
9. Utilization of Analysis Results


Business case
Data Collection
HDFS
Hadoop YARN
Data Modeling
Hadoop MapReduce
Data Processing
Apache Spark
HIVE
Data Visualization