Milestone 3: Design Document

Abstract

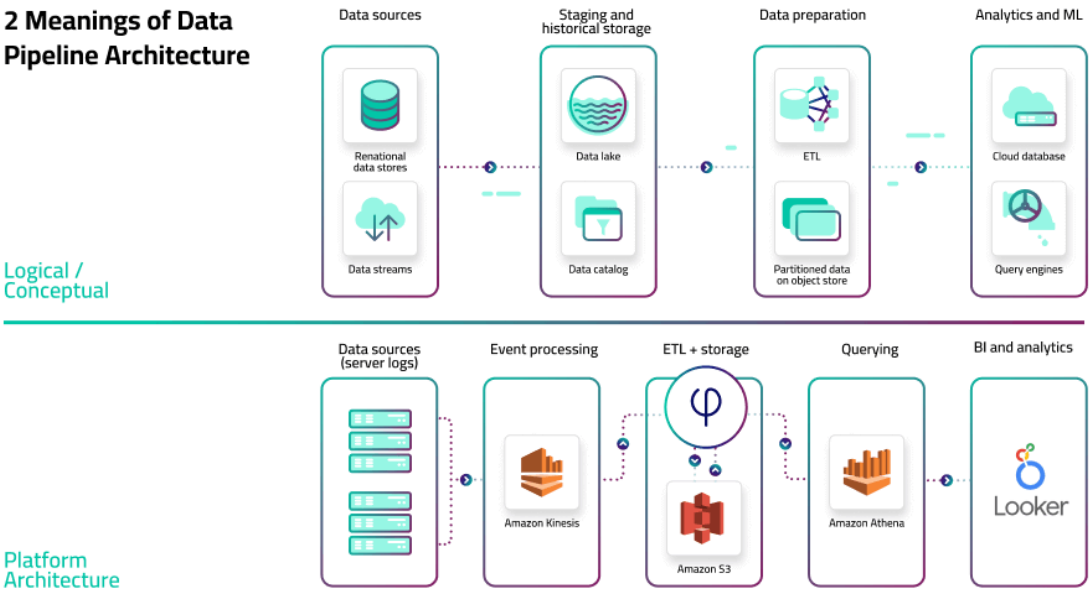
This design document outlines the architecture and components of the "Baseball Analytics" project, which aims to harness big data from MLB and other sources to provide advanced baseball analytics. The integration of Kafka, AWS, and FastAPI will empower teams, coaches, and fans with real-time and historical insights into the game of baseball.

Project Design

The project comprises several key components and design patterns to address the challenges of capturing, storing, processing, and analyzing big data.

Component Overview

1. **MLB Data Collection:** This component gathers extensive play-by-play datasets, player statistics, historical game outcomes, and predictive metrics. Real-time game data monitoring enables timely analytics and predictions.
2. **Data Storage & Processing:** Diverse datasets will be organized and stored efficiently, ensuring optimal querying capabilities. Real-time and historical data will be processed to derive actionable insights.
3. **Analytics & Data Retrieval:** This component provides tools for teams and coaches to analyze player performance, strategize game plans, and scout opponents. It also offers fans in-depth statistics to enhance their game-watching experience.

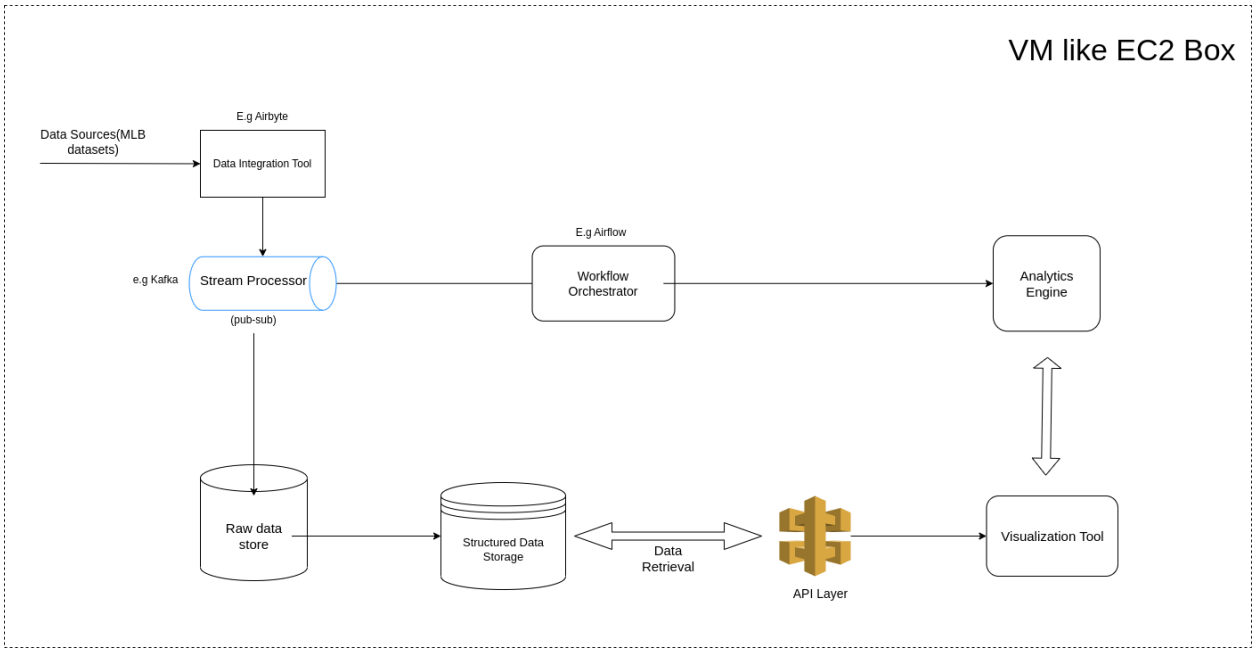


Data Flow

The flow of data in the system includes data collection from various sources, storage in AWS, and processing using FastAPI for real-time and historical analytics.

Architecture

The architecture of the system consists of AWS for data storage, Kafka for real-time data streaming, and FastAPI for data processing and retrieval. This integration ensures scalability and efficiency.



Interaction Diagrams

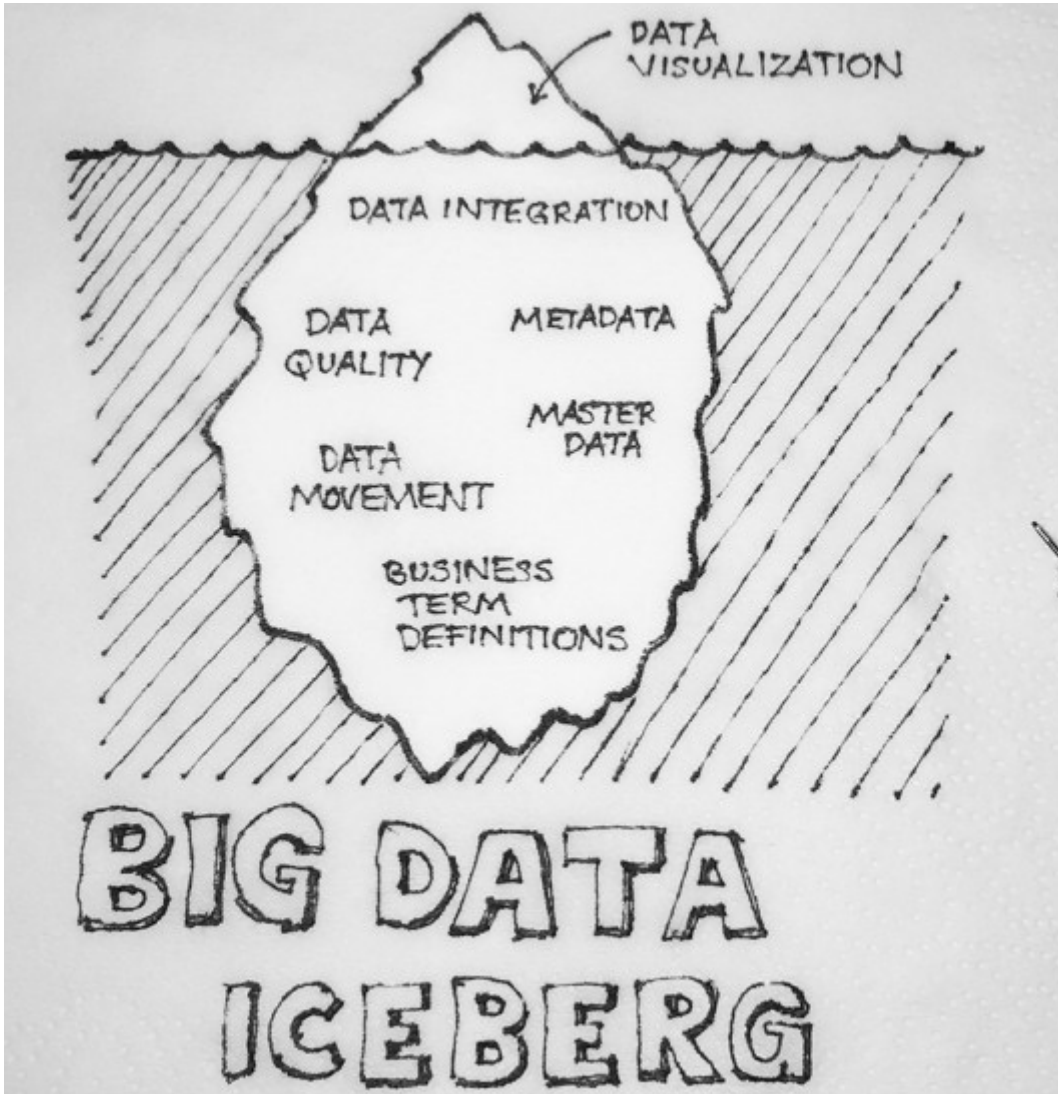
- **Use Case Diagram:** An overview of system functionality, including data collection, storage, and retrieval for teams, coaches, and fans.
- **Deployment Diagram:** Illustrates the software and hardware components involved, including AWS infrastructure, Kafka clusters, and FastAPI servers.
- **Sequence Diagram:** Demonstrates the sequence of actions for data processing, from data collection to analytics generation.

Big Data Challenges

The project addresses core big data challenges as follows:

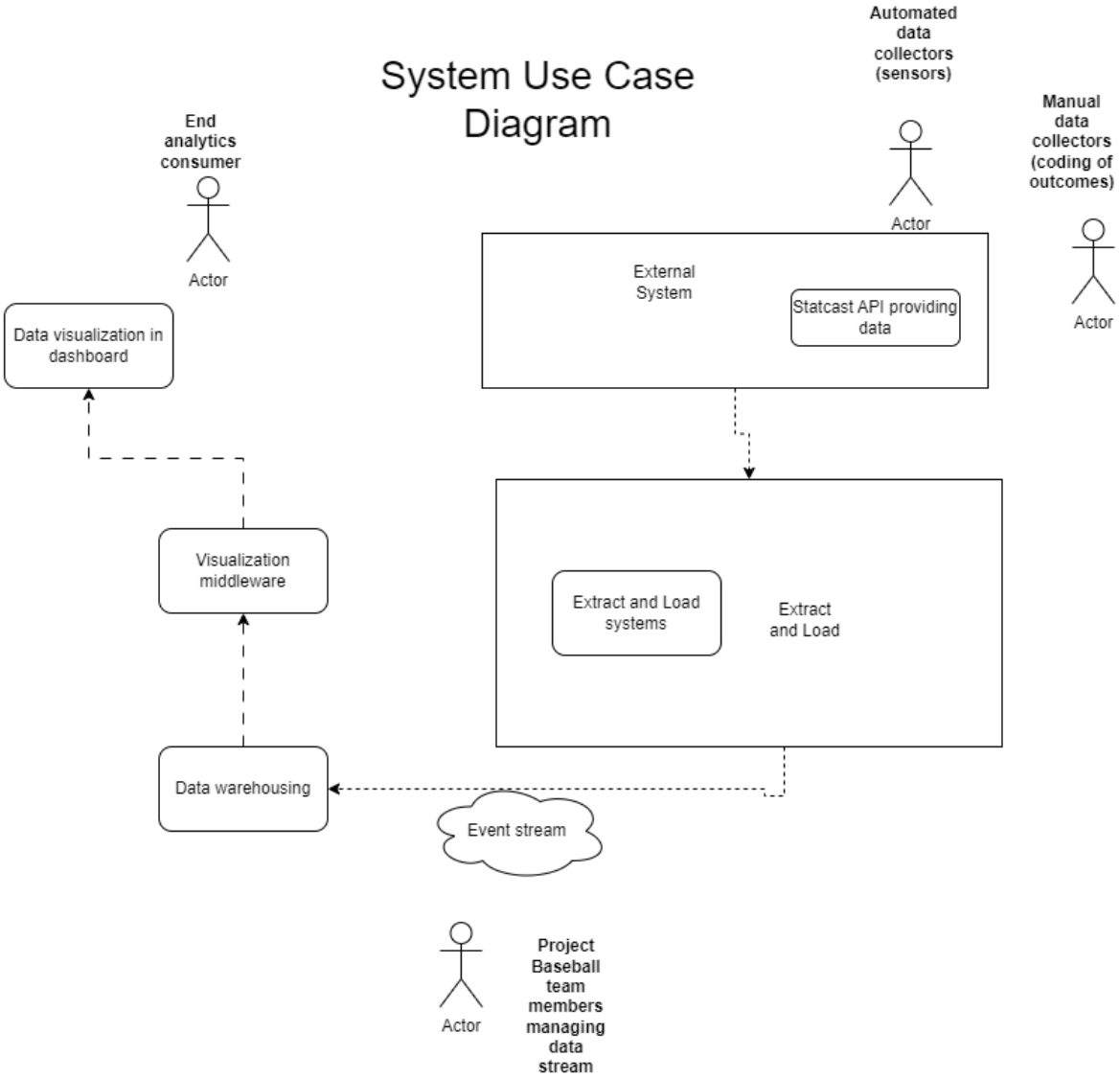
- **Data Authenticity:** Rigorous validation to ensure data authenticity.
- **Data Volume vs. Value:** Focus on extracting the most insightful data.

- **Consistency & Standardization:** Standardizing metrics and definitions for consistent analysis.
- **Accessibility & Restrictions:** Overcoming barriers to comprehensive data collection.
- **Integration of Historical Data:** Merging historical records for trend analysis.

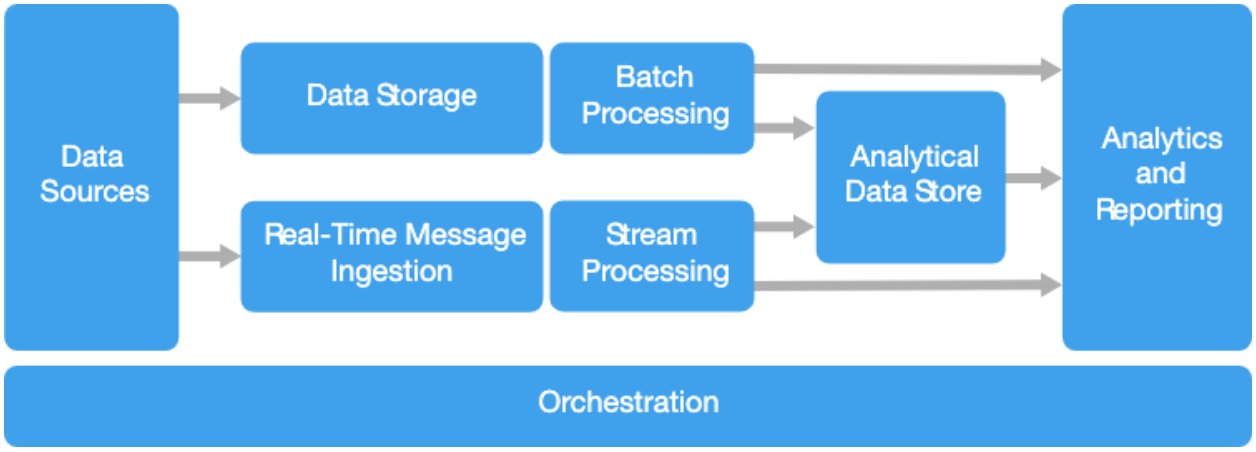


Design Diagrams

Use Case Diagram



Sequence Diagram



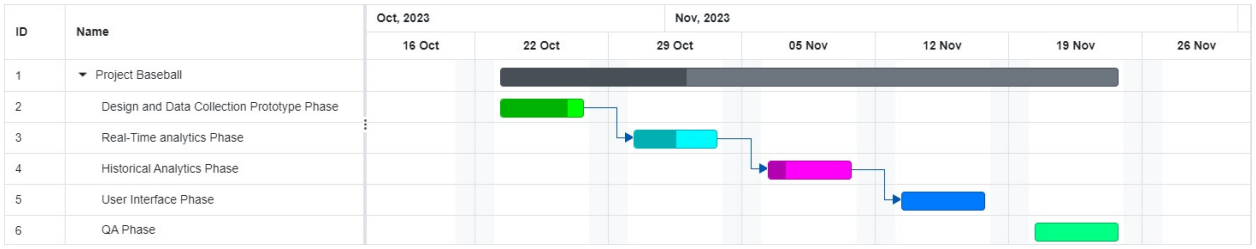
Project Schedule

The project will be executed in several prototypes, each adding more functionality and complexity to the system.

Prototypes

1. **Data Collection Prototype:** Collect MLB data, perform basic data processing, and store it in AWS. (Responsibility: Data Engineers)
2. **Real-time Analytics Prototype:** Set up Kafka for real-time data streaming and FastAPI for real-time analytics. (Responsibility: System Architects)
3. **Historical Analytics Prototype:** Integrate historical data and enhance analytics capabilities. (Responsibility: Data Scientists)
4. **User Interface Prototype:** Develop user interfaces for teams, coaches, and fans. (Responsibility: Web Developers)
5. **Testing and Evaluation:** Test the system for correctness and performance. (Responsibility: Quality Assurance)

Timeline (Gantt Chart)



Responsibilities

- **Data Engineers:** Responsible for data collection and basic processing.
- **System Architects:** Set up Kafka and FastAPI for real-time analytics.
- **Data Scientists:** Handle historical data integration and advanced analytics.
- **Web Developers:** Create user interfaces for teams, coaches, and fans.
- **Quality Assurance:** Ensure the correctness and performance of the system.