Big Data Computing & Analysis

Sprint 1

Haoyu Xu, Yang Qiao

Sprint 1 Schedule

- Define user stories of the Big Data Analysis

- Determine the technology used in the project

- Design the modular architecture for data input

- Design the system infrastructure of the project

User Stories

For mobile operators(AT&T, T-mobile):
- Collect and analyze users' raw data for personalized services such as family package
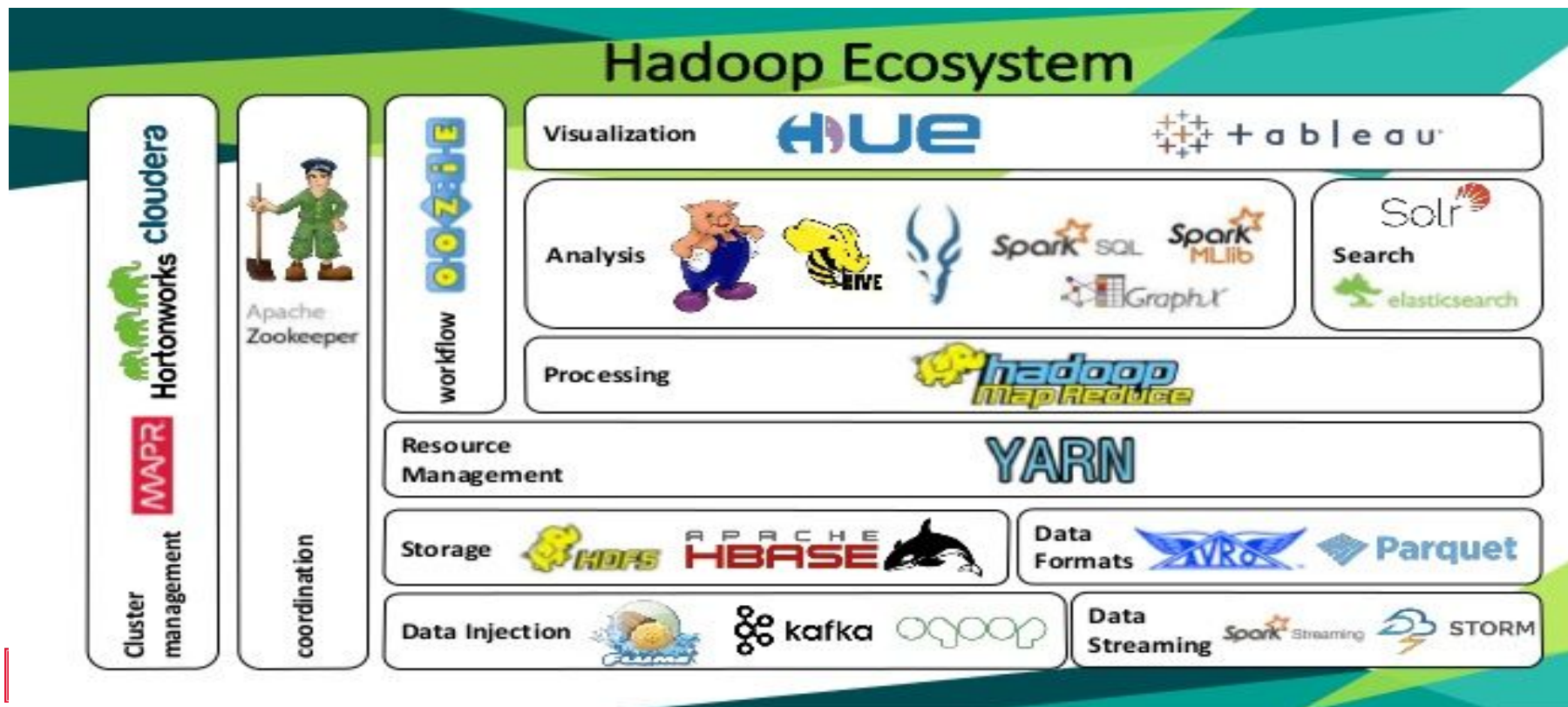- Visualize user data in the website for prediction

For mobile customers:
- Get call logs and other user information like intimacy
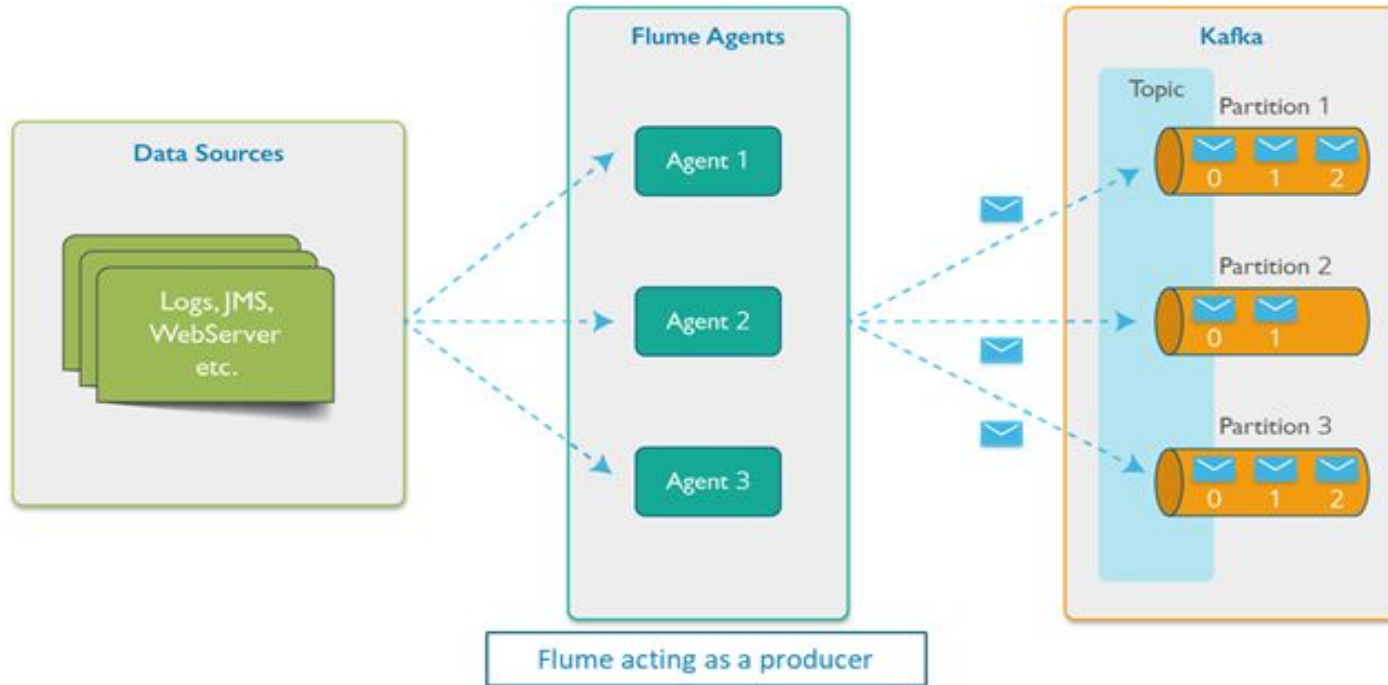- Reschedule service and data plan accordingly

Technology Used

- Apache Flume: Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows.

- Apache Kafka: Kafka is a distributed streaming platform. We use it to build real-time streaming data pipelines that reliably get data between systems or applications.

Hadoop Ecosystem

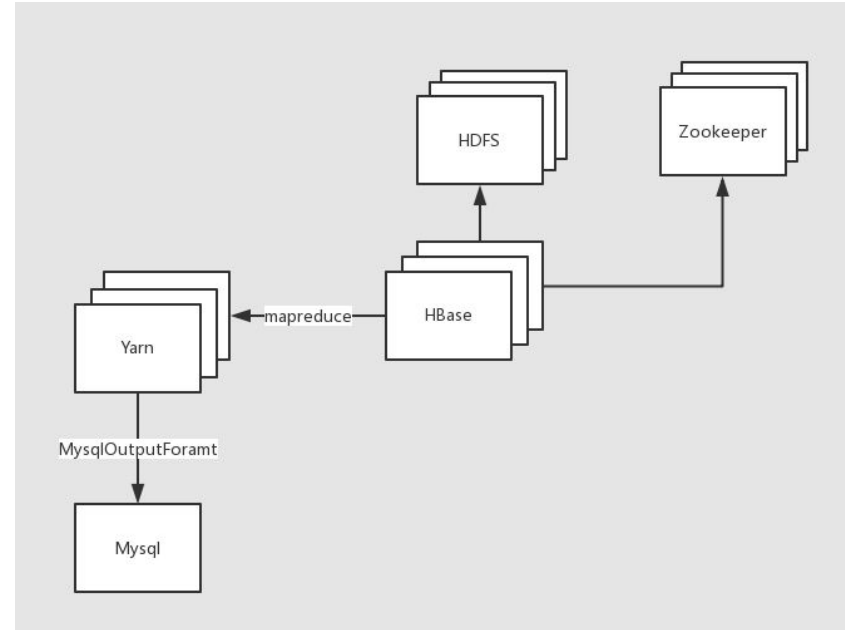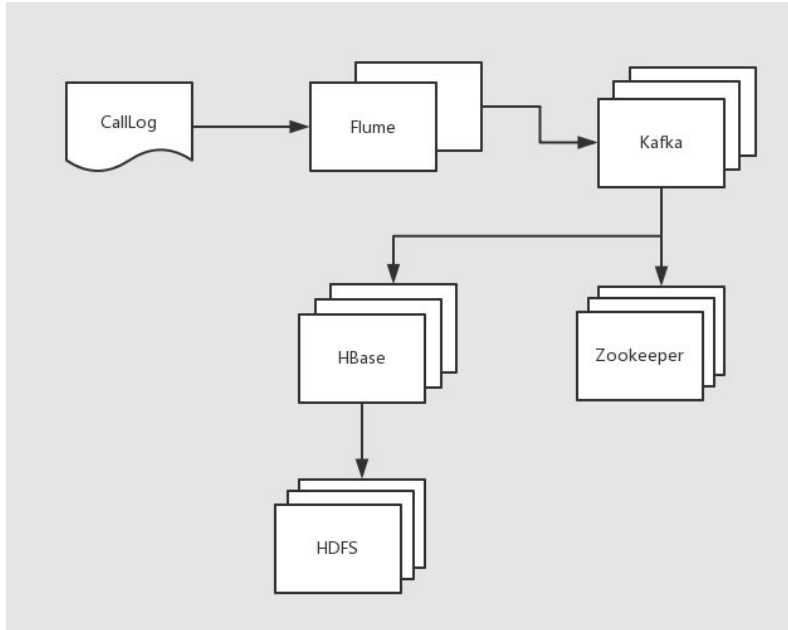# Modular architecture in Sprint 1

APIs

- Apache Kafka: *https://mvnrepository.com/artifact/org.apache.kafka*

- Apache Flume: http://www.apache.org/dyn/closer.lua/flume

- Apache Hbase: *https://mvnrepository.com/artifact/org.apache.hbase*

System Infrastructure

Task assignments

- Haoyu Xu: study on the implementation of Kafka

- Yang Qiao: study on the implementation of Flume

- Both engaged in designing system infrastructure and deciding technical stack