RBDA Proposal Submission

Social Media Technical Topics Analysis System - Project Proposal

Team Member:

1. Zhenghan Nan (NetId:zn2145)
2. Baijia Ye (NetId:by2352)
3. Junhao Fu (NetId:jf4519)
4. Yujia Zhue (NetId:yz10317)

1. Data Sources and Team Assignments

Member 1(Zhenghan Nan): StackOverflow Dataset

- Source: Stack Exchange Data Dump
- Download Link: https://archive.org/details/stackexchange
- Size: 5GB
- Format: XML
- Update Frequency: Quarterly
- Content: Technical Q&A, tags, user information
- Responsibility: Data profiling and cleaning, classification baseline construction

Member 2(Junhao Fu): Reddit Dataset

- Source: Pushshift Reddit Dataset
- Download Link: https://files.pushshift.io/reddit/
- Size: 3GB
- Format: JSON
- Update Frequency: Monthly
- Content: Posts and comments from tech-related subreddits

- Responsibility: Data profiling and cleaning, community analysis

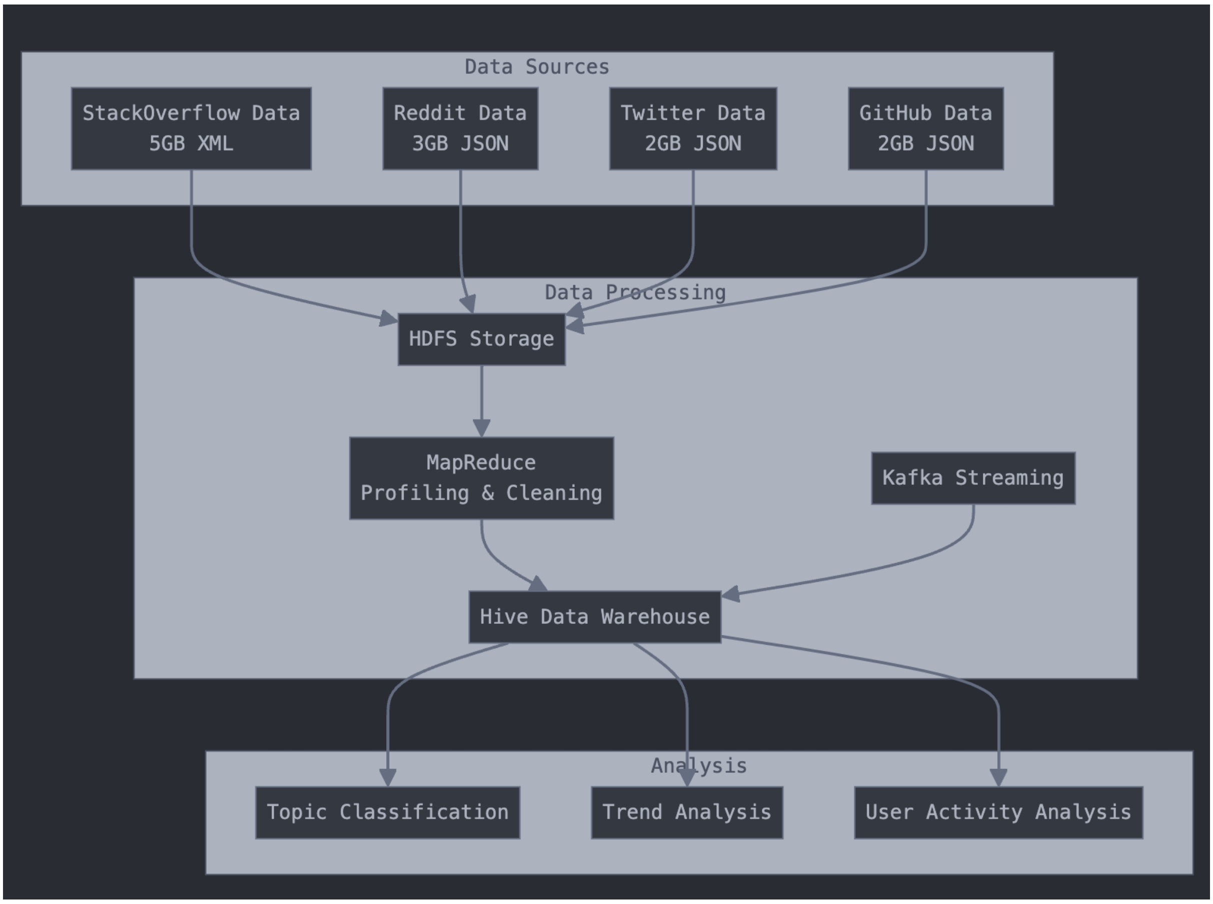
Member 3(Baijia Ye): Twitter Dataset

- Source: Twitter Developer Archive
- Download Link: https://developer.x.com/en/docs/tutorials/choosing-historical-api
- Size: 2GB
- Format: JSON
- Update Frequency: Real-time stream
- Content: Tweets with tech-related hashtags
- Responsibility: Data profiling and cleaning, topic tracking

Member 4(Yujia Zhu): GitHub Dataset

- Source: GH Archive
- Download Link: https://www.gharchive.org/
- Size: 2GB
- Format: JSON
- Update Frequency: Hourly
- Content: Issues, PRs, Commits data
- Responsibility: Data profiling and cleaning, developer behavior analysis

2. System Architecture Design



Storage Layer

- HDFS: Raw data storage
- Hive: Data warehouse for query and analysis

Processing Layer

- MapReduce: Data profiling and cleaning
- Kafka: Real-time data stream processing

Analysis Layer

- Topic classification
- Trend analysis
- User behavior analysis

3. Data Processing Flow

Data Profiling

1. Data completeness check
2. Data type analysis
3. Value distribution statistics
4. Missing value detection

Data Cleaning

1. Format standardization
2. Missing value handling
3. Outlier processing
4. Duplicate data removal

Data Integration

- Unified Hive table management
- Cross-source data association
- Unified query interface

4. Analysis Objectives

Topic Analysis

- Technical topic identification
- Hot topic tracking
- Topic evolution analysis

User Behavior

- Activity level analysis
- Participation pattern study
- Influence assessment

Platform Comparison

- Topic distribution differences
- User group characteristics

- Content propagation patterns