A Hadoop-driven Data Analysis System

Ashish Jindal
Rutgers University
Piscataway, NJ, USA
Email: ashish.jindal@rutgers.edu

Yikun Xian
Rutgers University
Piscataway, NJ, USA
Email: siriusxyk@gmail.com

Sanjivi Muttena
Rutgers University
Piscataway, NJ, USA
Email: sanjivi.muttena@rutgers.edu

***Abstract*—**The complexity of modern analytics needs is outstripping the available computing power of legacy systems. Distributed system like Hadoop compliments this requirement for storing and analyzing huge sets of information by providing a platform for parallel processing of large data sets stored over multiple machines. This project aims to setup a Hadoop infrastructure and demonstrate its power for big data analysis. Some machine learning models will also be deployed in this system so that developers can directly call corresponding APIs to execute training and testing models.

***Index Terms*—**Hadoop-driven, Data Analysis, Infrastructure

I. Project Description

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Hadoop stores data as it comes in - structured or unstructured - saving time on configuring data for relational databases. In our project we will store a large dataset in Hadoop file system and use it for analysis of events from various sources like news, twitter etc. Our project falls in the category *Scalable Algorithms Infrastructure*. The main stumbling points for this project are identifying a large dataset for Hadoop, configuring Hadoop for multiple clusters and building a data analytics system over Hadoop.

The project has four stages: requirement gathering, design, infrastructure implementation, and user interface.

A. Stage1 - The Requirement Gathering Stage

Following are the deliverables for this stage:

a) General Description: This project is about setting up an infrastructure for big data analytics using Hadoop and demonstrate a data analytics application using event data from various sources like news and blogs, etc. The first part is to construct a distributed computing platform based on Hadoop and its high-level applications like Mahout. The system provides some well-encapsulated APIs for training and testing general models. The second part include a real application on event analysis driven by Hadoop. It mainly demonstrate the feasibility of architecture to apply Hadoop into existing web-based application.

b) User Type 1: People interested in analyzing data that can't be stored on a single machine.

- User Interaction Modes: Hive query language.
- Real World Scenarios:
  - Scenario 1 Description: Processing of large server logs.
  - System Data Input for Scenario 1: Log files.
  - Input Data Types for Scenario 1: Structured log files.
  - System Data Output for Scenario 1: Information based on query.
  - Scenario 2 Description: Text mining.
  - System Data Input for Scenario 2: Big sources of textual information.
  - Input Data Types for Scenario 2: Text.
  - System Data Output for Scenario 2: Information based on query.

c) Project Timeline and Division of Labor: We will start with studying about Hadoop and map reduce systems and then set up a Hadoop system on local host. When we are successful in doing that we will find a small dataset, large enough to fit on a single system and start test analysis over it using Hadoop. Then we will repeat the same process by setting up Hadoop in pseudo distributed mode. Parallely two team members will start building the demo analytics application over Hadoop. In the end we will setup Hadoop on multiple clusters and deploy our application over it.

- Week 1 and Week 2: Work on requirement gathering and design of system using Hadoop. Setup Hadoop on local host as a single cluster and perform dummy analysis to test its functionality.
- Week 3 and week 4: Setup Hadoop infrastructure over multiple clusters and create an analytics application demonstrating big data analysis using Hadoop.
- Week 5: Test the application and prepare the necessary documentation & project report.

References

[1] M. Naughton, N. Kushmerick, and J. Carthy, "Event extraction from heterogeneous news sources," in *proceedings of the AAAI workshop event extraction and synthesis*, 2006, pp. 1–6.

[2] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[4] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[5] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*. IEEE, 2010, pp. 1–10.

[6] T. White, *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.