BITS Pilani
Pilani | Dubai | Goa | Hyderabad

BIG DATA SYSTEMS – ASSIGNMENT 2

Submission Date: 20 Dec 11.50 PM

Weightage: 15%

1. Business Context

BITS Pilani is conducting the examinations for the Work Integrated Learning Programme Division. The job of conducting the examinations is handed over to the third party vendor who is conducting the examinations using the infrastructure that is hosted within its data center. The complete exam data (students, exam, evaluation and logs etc.) resides within the vendor's data storage. Based upon the requirements of BITS Pilani, the appropriate reports are made available for the analysis purposes. But this is not adequate for the ad-hoc analysis that needs to be carried out on this exam data to check how the exams are going on, what are the major issues those are surfacing out during exam conduction etc.

2. Business Problem Understanding

While online proctored examination is going lot many interesting facts are getting captured about the candidate, proctors and examinations etc. in real time manner. All these details are stored in a raw format as logs in storage. Many times students faces issues during the exam like internet connectivity issues, power issues etc. which can cause effects on the durations of the exam. Some time they are not able to upload the images or uploaded images are not visible after upload. When their live feed is captured, the system also detects various things face not visible, audio in the room, deviation to another window etc. Sometimes questions or part of question is not clear or visible which might have been cleared by Proctor using the live chat option. There can be some instructions passed on from Proctor to students. These are just some the example scenarios that one can envision or encounter during the online exam.

The detailed logs for all these online exams are captured but it's not a simple task to go ahead and analyze these logs by human being as they are captured at very high rate like every 10 seconds and that too for huge number of students at a time like 5000 to 6000. But still the analysis of this raw data will surface out the commonly recurring problems which sometimes surfaces as the bottleneck and results into the whole process of examinations. BITS management is interested to check whether we can build up an analytics platform that can help in automating this analysis process which can span across different facets of examination like question paper setup process, online exam conduction, evaluation and reevaluation process. This will give them the better vision into the set of problems that surfaces out and enable to resolve them well ahead of time.

Some of the ad-hoc queries (but not limited to) for which they are seeking the answers can be summarized as follows:

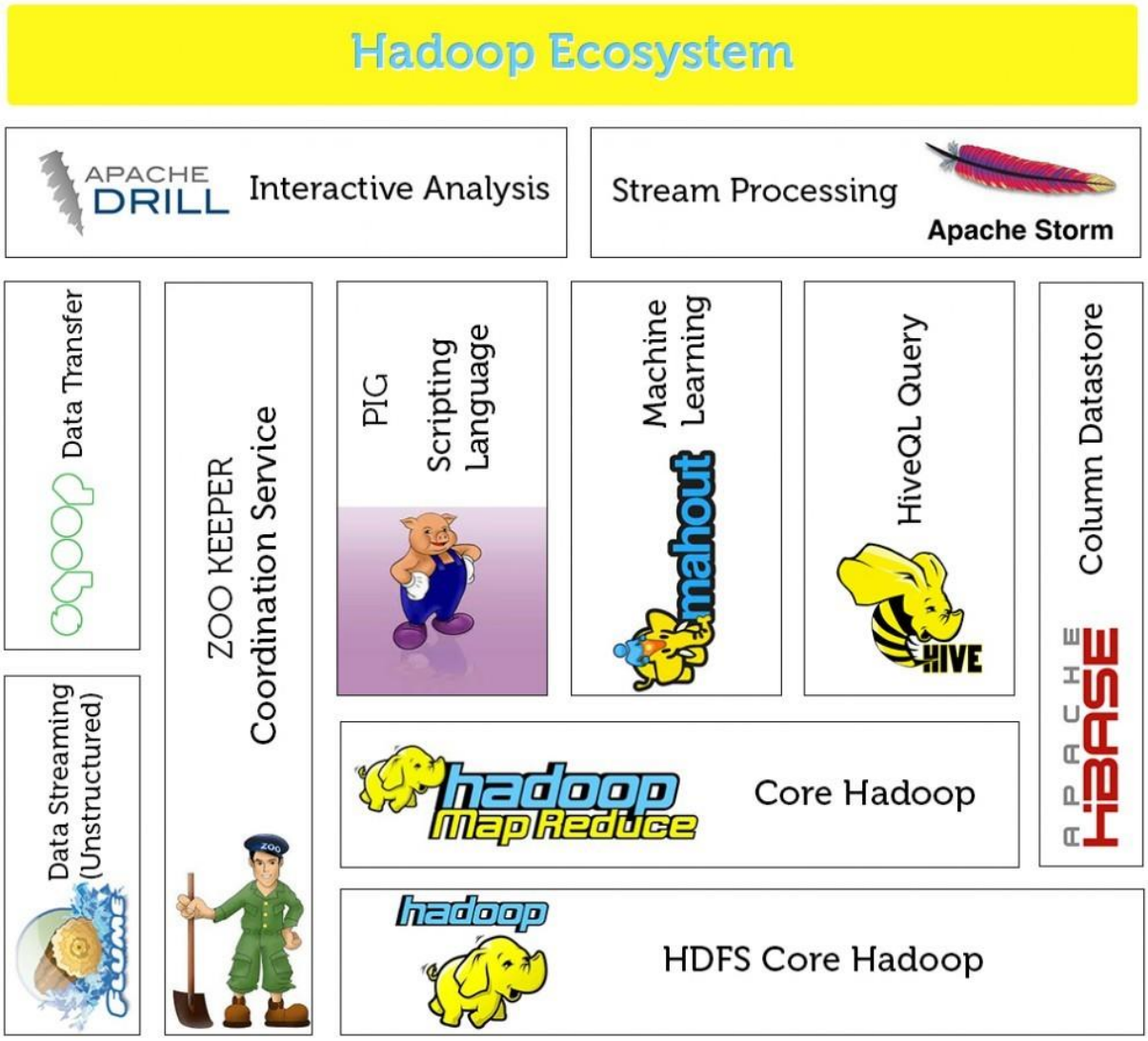
- What are the total number of issues raised during a particular exam session and their breakup by issue types?
- What is the trend of the issues for the past "n" online exam sessions?
- How much time on average students / tech team has to spend on resolving issues during a particular exam session? Trend of this measure over past "n" sessions.
- Is there any particular student/s who is facing common set of issues during all the exam session he /she is attending? What sort of issues are being faced by him or her.
- What is the average time spent by the proctors over the online chat with candidates during exam sessions?

3. Data and design Considerations

BITS management has given a go-ahead for a prototype that will enable the analysis of raw data that is captured at time of actual exam time. Usually this exam event data resides into the vendor's storage system and contains details of various events as described earlier for all the students who are appearing for the examinations. You can assume that even if student is not facing any issue while writing the exam still a log entry is made for him every 10 seconds. Also based on your experience as a student you are free to make assumptions about the raw data format that is being collected.

The high level idea is to collect, process and analyze the online exam events raw data using Apache Hadoop and related ecosystem projects. This high level expectation can be broken down into sections like storage, ingestion, processing, analyzing and orchestration.

- Storage refers to decisions around the storage system (HDFS or HBase), data formats, and data models
- Ingestion refers to getting the raw data from the vendors system and secondary data like students profile from BITS systems and loading it into Hadoop for processing
- Processing refers to taking the raw data ingested into Hadoop and transforming it into datasets that can be used for analysis and querying.
- Analyzing refers to running various analytical queries on the processed data sets to find out answers and insights to the questions prescribed in earlier section
- Orchestration refers to automating the arrangement and coordination of various processes that perform ingestion, processing and analyzing



4. The Data Engineering teams Responsibility

As a part of Data Engineering team, your group's responsibility is to check the feasibility of this proposed architecture and build a working prototype that will help in addressing the stressing queries.

For that purpose, you need to

1) Propose a high level system architecture that will enable the end to end data flow for this proposal
2) Identify candidate tools from the Hadoop ecosystem or otherwise that will be suitable for the particular task and decide upon which one can be leveraged citing out the reasoning for the selection

3) Build a data pipeline that will enable the end to end flow of data across various system components identified earlier
4) Provide a simple query interface through which the answers to the sample questions can be obtained by naïve users
5) Build a reporting component that will periodically generate a detailed report providing the answers to the all queries listed by the management

5. Submission Requirements

As a result of this you need do three types of files

1) Project report document
2) Python Code / queries / other relevant code wherever it makes senses
3) A video demonstration of design and working prototype

You can check for the following aspects (but not necessarily restricted to) while completing the submission requirements

a) Architecture diagram with appropriate description of the components / tools with reasons for their selection, communication between them
b) Necessary assumptions made before implementations of the project with appropriate descriptions for the same
c) Critical decisions needs to be made for data schema / model , data formats or any other important aspect related to the data to be processed
d) Configurations / settings done for the ecosystem components and reasoning for the same
e) Data flow between the components in the system
f) Data processing steps with a quick summary of requirement of each step, data inputs and expected outputs from them
g) Ad-hoc queries analysis , input query formats , expected outcomes
h) Working of periodic report generation module
i) Automation of the entire / part of the scenario
j) Any other consideration that you feel appropriate to explain your project in better manner

General Notes:

- Refer the document used while registering the groups. In case of discrepancies, write to me separately (copying all your group members) with subject line as "DSE BDS Group <your_group_number>". email - ppawar@wilp.bits-pilani.ac.in
- Using the Canvas, only the first member of group (as listed in the above mentioned doc) has to upload the file. No submission over email will be considered.

- Name the document file in format like "Grp_<your_group_number>.zip" only. Don't add anything into the file names.
- Use this Google drive location for sharing the demo videos. Follow the same naming conventions for them.
- Make sure that you upload the file well ahead of deadline. At last moments, we have seen several groups have faced issues while doing the submissions.
- Note - As it's a group assignment, only one submission is expected from each group. Unnecessary don't upload the solution on individual basis. If it's observed, then the penalty (25% reduction) will be applicable on it.

6. References

- Hadoop Application Architectures