DB Management Systems: Course Introduction

Joel Klein – jdk514@gwmail.gwu.edu

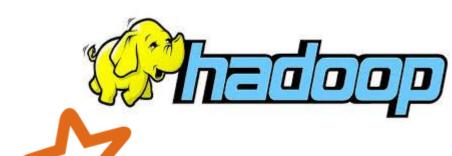
The Necessity of This Course

The Problem of "Big Data"

- "Every 60 seconds on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded." https://zephoria.com/top-15-valuable-facebook-statistics/
- This volume of data alone is enough to break traditional DBMS (Database Management System[s]), but this is only compounded by the fact that most of this data is not in a structured format
 - Traditional DBMS are focused on structured data
- This is the primary driver behind NoSQL databases
 - NoSQL literally just means that it is NOT SQL
 - Not being SQL just means that they don't use tabular relations

Why do NoSQL DBMS Solve "Big Data"

- NoSQL DBMS help diminish some of the problems presented by "Big Data" via a couple techniques:
 - Scalable Infastructure
 - Sharding/Replication/Redundancy
 - Distributed Computation
 - Data Structure
 - Key-Value
 - Document
 - Etc.



mongoDB_®

So Bye-Bye SQL?

- SQL and RDBMS are still very important in data processing
- RDBMS still provide a lot of guarantees in the manner in which data is handled and processed
- SQL also remains a very powerful method to query information
 - Which is why many DBMS try to emulate SQL and create SQL-like query languages
 - SQL also enables easier optimization for querying

Analytics in "Big Data"

Big Data Analytics

- This explosion of data generation has required an equivalent response on the analytics front
- Luckily, the scalability and flexibility of these new DBMS systems can also lend itself to the analysis of said data
 - With distributed systems, we can more effectively query data
 - With unstructured and non tabular data, we can be more inventive with our queries
- In addition, systems like Hadoop and Spark present means for distributed processing of large data systems

Putting It All Together

Why This Course Exists

- Data analytics/Machine Learning/Data Science are nothing without data
- This course is focused on how and why we store and access the data that fuels our analysis
 - This is important as many of todays problems cannot be handled by csv files and a singular computer
- So, we will focus on how to store what data where
 - Using the more dominant technologies currently in the industry

End Slide

EMSE 6586 – DBMS for Data Analytics