CSP554
BIG DATA TECHNOLOGIES

End of Term Ideas

Research Paper and Projects

- No final in this course
- Paper or project due the Wednesday of finals week
- Provides an opportunity for researching or exploring big data technology in more depth
- Three phases:
  - ½ paper proposal with 2-3 citations
  - 3-4 page draft mainly covering review of the literature, not code or other project deliverables
  - Final paper or project
- I am easy going about due dates except for the final project as grades must be provided on time

Research Paper Details

- Single student
  - Since you don’t need to collaborate with others, I am a bit more “picky” about your work
- 12-14 pages (whatever format you like)
- Must have references and citations
- In depth description, analysis and interpretation of some topic based on a search of the literature

Project Details

- Group of 3-4 students (as selected by professor)
- 10-12 pages (whatever format you like)
- Must have references and citations
- Must include description of what work each student contributed at the start of the paper in a special section titled “Contributions”
- 1/3: review of literature to describe the technology you are using in detail
- 1/3: detailed description of what experiment or demo you are doing along with expected results and objectives
- 1/3: Results and their analysis with charts and graphs as needed
- Appendix: example data and code or other similar materials

Indicating Your Interest

Project or Paper

- Join one of the groups created to allow you to indicate your interest
  - Complete by Monday 10/1 – No exceptions
  - To join a group select the “Project/Paper Groups” selection form the Blackboard “Assignments & Project” page
  - You can only join one group, so think about what you might want to do generally before making you selection
  - If you make a mistake/typo in your selection, don’t worry
    - Send an email to me with your correct choice and I will make adjustments (one way or another)

Project Mechanics

- I will form groups of 3-4 students (not your choice)
- I will notify you about your group membership and share your emails
- You should reach out to each other to decide on the exact details of your project proposal
- All of you will collaborate together to write the ½ page project proposal
- One student in each group, which I will designate, will be responsible for uploading the proposal, draft and final project
- All students in a group will be evaluated for their contributions individually and together
- Each student will receive a separate grade for each project phase
- If at any point you have concerns with members of your group try to work together to resolve it (this is what happens in the "real world")
- If that does not lead to success, feel free to reach out to me any time
  - All such communications will be keep confidential

Example Research Paper Topics

- Compare and evaluate
    - Other execution engines: Apache Beam, Apache Flink
    - Other Big Data SQL engines: Impala, Drill, Presto, AWS Athena, Microsoft Azure U-SQL
    - Compare and evaluate cloud NoSQL databases: AWS DynamoDB and Microsoft Azure CosmosDB and others
    - AWS Kinesis vs Apache Kafka
- Research Hadoop Security: Sentry, Ranger, Knox, latest other papers
- Research NoSQL Database Security
    - Compare and evaluate security from among Cassandra, HBASE, Accumulo, MongoDB
    - Compare and contracts the security of some cloud NoSQL databases such as AWS DynamoDB, AWS DocumentDB, and Microsoft Azure CosmosDB and others

Example Research Paper Topics

- Review Big Data Graph Processing: Neo4j database, Gigraph, Spark GraphX, others
- Explore and evaluate big data machine learning Spark MlLlib, H2O, SparkR, TensorFlow, others
- Compare and evaluate different (near) real-time big data processing systems: Spark Streaming, Storm, Kinesis, others
- Review and compare NewSQL Database architectures and functionality
  - Redshift, SQL Data Warehouse, Snowflake, MemSQL, Splice Machine
- Explore the use of Big Data Technologies in Industry
  - Healthcare, Finance, Ecommerce, Logistics, etc.
- Explore data search technologies:
  - SOLR, ElasticSearch

Example Projects

- Implement example applications in several tools to explore their differences and capabilities in practice
  - Big data machine learning tools (Tensorflow, Sagemaker, Spark Mllib)
  - Big data graph processing tools (Graphx, Neo4J)
  - Big data SQL tools (Athena, Presto, Drill)
  - Real time big data processing (Kafka, Kinesis)
  - Other NoSQL databases (DocumentDB,, CosmosDB, DynamoDB, CouchDB)
  - Other Big Data Exec
- Explore the use of text search tools: SOLR, Lucene, Elastic Search, cloud alternatives
- Apply big data techniques to explore some interesting data and derive insights
- Create big data processing pipelines using a combination of tools such as:
  - Kafka, spark streaming or storm or AWS kinesis, HDFS, and some other store such as HBASE or NoSQL database

Example Projects

- Explore the use of big data cloud technology
  - AWS EMR, AWS Kinesis, Azure HDInsight
  - AWS DynamoDB, AWS DocumentDB
  - Azure CosmosDB
- Create a system that accepts a domain specific language something like Pig Latin and outputs Spark code
- Look to Kaggle for interesting data sets
  - https://www.kaggle.com/datasets
  - You don’t need a huge data set to demonstrate use of big data technologies

Example Projects

- Implement example applications in several tools to explore their differences and capabilities in practice
  - Big data machine learning tools
  - Big data graph processing tools
  - Big data SQL tools
  - Big data graph processing tools
- Explore the use of text search tools: SOLR, Lucene, Elastic Search, cloud alternatives
- Apply big data techniques to explore some interesting data and derive insights
- Create big data processing pipelines using a combination of tools such as Kafka, spark streaming or storm, HDFS, and some other store such as HBASE or NoSQL database

Example Projects

- Evaluate the performance of multiple big data machine learning tools
  - Spark MLlib, H2O, SparkR, TensorFlow, others
- Compare and evaluate
  - Other execution engines: Apache Beam, Apache Flink, AWS Glue
  - Other Big Data SQL engines: Impala, Drill, Presto, AWS Athena, Microsoft Azure U-SQL
  - Compare and evaluate cloud NoSQL databases: AWS DynamoDB and Microsoft Azure CosmosDB and others
- Test Big Data Graph Processing: Neo4j database, Gigraph, Spark GraphX, others
- Benchmark the performance of different (near) real-time big data processing systems: Spark Streaming, Storm, Apache Flink, Apache Beam, others

Research Paper or Project Draft

- 3-4 page draft mainly covering review of the literature
- Must include 2-4 PROPERLY formatted citations from your review of the literature
- For project

  Must include a list of 5-8 main milestones for the project, the due dates and the owners