CSE5BDC Exam Revision Part 2

We will go over some of the main topics and concepts covered in this subject to help you prepare for the exam. Remember the emphasis for the exam is to test your understanding. So, just remembering a lot of facts will not help much by itself. You need to also understand how the concepts are connected and the implications of using the various technologies that have been taught.

We now outline the important topics to study grouped by the week it was taught. Please study the topics below and answer any of the questions asked below. Please then attend the face-to-face exam revision lecture and ask the lecturer about any of the topics you are unsure about.

**Week 1**

1. Know the motivation of cloud computing well.
2. Differentiate between:
   1. infrastructure as a service
   2. platform as a service
   3. software as a service.
3. Why are virtual machines so important for the cloud?
4. What is the difference between S3 and EBS?
5. Why study Big Data?
   1. Examples of where we can find Big Data.
6. Given a particular problem, should we use MapReduce to solve it or not?Why?
7. Know the nine big concepts of MapReduce.
8. Know how mappers and reducers work.
9. Know what happens to the data when it leaves the mapper and arrives at the reducer.
10. Know the difference between in mapper aggregation and combiner.

**Week 2**

1. Know how HDFS works.
2. Know how jobs are scheduled and executed in the MapReduce framework.
   1. Job tracker, task tracker, etc.
3. How does Hadoop achieve fault tolerance?
4. How does Hive differ from a traditional relational database?
5. What is the relationship between Hive and MapReduce?
6. How can we write Hive queries that executes faster?

**Week 3**

1. What are the benefits of using YARN?
2. How does YARN achieve these benefits?
3. Why is MapReduce ill-suited for processing streaming data?
   1. How does Storm address these limitations?
4. Why is graph processing so important?
5. Why MapReduce is so bad for graph processing?
   1. How does Pregel address these limitations?
6. What are the advantages of Spark over MapReduce?
7. How does Spark handle fault tolerance?
8. What are RDDs?
9. Why does Spark perform lazy execution?
10. How does Scala differ from Java?
11. Why did the creators of Spark decide to build Spark on top of Scala?

**Week 4**

1. How does Spark schedule jobs?
2. What is the difference between narrow and wide dependencies?
3. Make sure you practice programming Spark RDD programs. You will be given problems and expected to write the solution in Spark. Real Spark RDD code not pseudo code.
4. Why use SparkSQL?
   1. ease of use
   2. speed
5. How does Project Tungsten help to improve the performance of SparkSQL?
6. What at the advantages of using column storage format?
7. Make sure you practice programming SparkSQL programs. You will be given problems and expected to write the solution in SparkSQL. Please write your answer using the DataFrames/Datasets API instead of SQL

**Week 5**

1. Give at least two different Spark Streaming use cases.
2. What are the advantages of using structured stream?
   1. ease of use
   2. speed
3. Know the difference between event time and processing time.
4. How does GraphFrames make graph processing easy?
5. How does NoSQL differ from MapReduce?
6. How does NoSQL differ from traditional relational databases?
7. Why do many NoSQL stores use eventual consistency?
8. Know the CAP theorem.
9. What is the main difference between HBase and DynamoDB?

**Week 6**

1. What services of AWS can be used to achieve automatic elasticity?
2. How are these services used together?
3. What are the benefits in using the relational database service versus setting up everything by yourself on EC2?
4. Why is high availability more desirable than disaster recovery?