Big Data Advanced Analytics – Cast Study

As a Big Data Analyst you are expected to build/test & support large-scale data processing systems and be an expert in data ingestion, modelling, processing & prediction solutions using AI techniques (Deep Learning/Machine Learning) to meet business needs of the organisation. You should be able to work with the latest Big Data, NoSQL and Advanced Analytics tools (R, Python, Spark MLlib, Apache Flink, Apache Kudu, etc.,) to deliver cutting edge solutions. You should be creative problem solvers, resourceful in getting things done and productive working independently or collaboratively and well versed in Agile/DevOps methodology and behaviours.

Use-Case: Airline on-time performance

Reference Link: http://stat-computing.org/dataexpo/2009/

Have you ever been stuck in an airport because your flight was delayed or cancelled and wondered if you could have predicted it if you'd had more data? This is your chance to find out.

The data:

The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed.

1 Source data details

Download the stats created for year 2008 & 2007.
http://stat-computing.org/dataexpo/2009/the-data.html

Supplemental Data: http://stat-computing.org/dataexpo/2009/supplemental-data.html

1.1 Develop a solution using R/Python/Spark MLlib tools and Big Data platform to answer the following questions. Demonstrate your expertise in each tools.

- Which carrier performs better?
- When is the best time of day/day of week/time of year to fly to minimise delays?
- Do older planes suffer more delays?
- Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?
- Create a model to predict flight delays
- How well does weather predict plane delays?

Note:

- Prepare a technical presentation along with Architecture Diagram/UML to explain the solution for the above case-study and share
- Share the code & configuration details
- Walk us through the solution, model, tools and code during assessment