Data Pipelines, Hackathon Brief

Hackathon Part A– Pipeline Design

Objectives

In this exercise you will:

- Design and document a sequence of operations to convert a pair of source files into an output file that can be used in subsequent analysis
- Implement your design
- Test your pipeline with an appropriate test deck and document your test process and results

Overview

You have been provided with two zipped files in the folder PartA_Data. The first contains data about the time of sunrise, sunset and the length of the day in Edinburgh. The second contains weather data from a weather station in Scotland.

Your brief is to estimate the temperature every minute for the year 2012. This will be used in a further analysis of operational data of electronic equipment in the vicinity.

You may use any combination of the technologies that you have been introduced to together with any other tools that you are already familiar with. This lab does not give you detailed instructions.

Useful information

For the purposes of your modelling, you may assume a linear trend between the lowest and highest temperatures each day.

1. The highest temperature will be part way through the later part of the day.
2. The lowest temperature will be shortly after dawn. Remember to include the assumptions you make as part of the documentation of your pipeline.
3. You will need to explain your pipeline, documentation and naming conventions to the instructor. Remember this needs to be in a written format. The reason for this may become apparent in the next part of the Hackathon.
4. Hint: Your final data set should contain 527040 data elements consisting of the time and temperature every minute from midnight on 31st December 2011 till 23:59 12 months later.

Hackathon Part B – Pipeline Design Karma

Objectives

In this exercise you will:

- Execute the documentation of a sequence of operations to convert a pair of source files into an output file that can be used in subsequent analysis
- The data sources may not precisely match the original sources the pipeline was designed to process
- You must model the new data and then test and document any adjustments that are required to successfully process the new data sources

Overview

You have been provided with two zipped files in the folder PartB_Data. The first contains data about the time of sunrise, sunset and the length of the day in London. The second contains weather data from a weather station in Hertfordshire.
Your brief is to estimate the temperature every minute for the year 2012. This will be used in a further analysis of operational data of electronic equipment in the vicinity. You will use the pipeline developed by one of your colleagues **NOT** your own.

Useful information

For the purposes of your modelling you may assume a linear trend between the lowest and highest temperatures each day.

1. This lab does not give you detailed instructions but your colleague will have done
2. Your colleagues documentation will include any assumptions they made as part of their pipeline design.
3. You will need to provide a written assessment of the pipeline, documentation and naming conventions to the instructor and your colleague. After all bad documentation…
4. Hint: Your final data set should contain 527040 data elements consisting of the time and temperature every minute from midnight on 31st December 2011 till 23:59 12 months later.

Hackathon Part C – Big Data model building

Objectives

In this exercise you will:

- Develop a model that explores the relationship between ambient temperature and equipment performance
- Determine if there is any relationship between this model and customer complaints
- Consider the impact global warming might have on equipment performance and customer complaints
- Write a presentation that explains the projects results

Overview

You have been given the equipment performance data for a telecommunications mast (#79645381) in the Strathspey area of Scotland.
It details the workloads being undertaken by each of the 10 transmitters on the mast and the speed at which their cooling fans are spinning.

- Each transmitter has 3 fans in an N+1 configuration
- Minimum fan speed is 120 RPM delivering 3 CFM
- Maximum fan speed is 800 RPM delivering 20 CFM
- Maximum number of connections per transmitter is 168
- Maximum data bandwidth per transmitter is 20Mhz
- Fan failure rate per minute is nominally 1 in 3,285,000 (approx. 50,000 MTTF)
- Each connection uses 1/4 W and each 1Mhz of data bandwidth uses 2W

You also have error log, service log and customer complaint data files.

Useful information

1. This lab does not give you detailed instructions
2. Develop a model that explores the relationship between ambient temperature and equipment performance. The final data set you generated in Part A may be useful.
3. Determine if there is any relationship between this model and customer complaints.
4. Your end-users are concerned that temperature rises due to Global Warming might have an adverse effect on the mast equipment and consequently on customer satisfaction. Consider what level(s) of temperature rise might have an impact on mast operations.
5. Prepare a presentation that explains your methodology to a lay audience and presents your conclusions.

You may share this presentation with your fellow students and your instructor.

Bonus question

What are the busiest times of day during the week and does it differ at weekends?

❑ Yes ❑ No

Does this match your expectation?

Hackathon Part D – Compare Big Data sets

Objectives

In this exercise you will:

- Compare multiple data sets with a previously developed model, modifying the model where necessary
- Determine if any of the data sets are unusual
- Report on your findings

Overview

You have been given the equipment performance data for eight telecommunications masts (odd numbers from #52862437 - #52862451) in the Royston area of Hertfordshire.
It details the workloads being undertaken by each of the 10 transmitters on each mast and the speed at which their cooling fans are spinning.

- Each transmitter has 3 fans in an N+1 configuration
- Minimum fan speed is 120 RPM delivering 3 CFM
- Maximum fan speed is 800 RPM delivering 20 CFM
- Maximum number of connections per transmitter is 168
- Maximum data bandwidth per transmitter is 20Mhz
- Fan failure rate per minute is nominally 1 in 3,285,000 (approx. 50,000 MTTF)
- Each connection uses 1/4 W and each 1Mhz of data bandwidth uses 2W

You do not have access to the error log, service log and customer complaint data files for this area.

Useful information

1. This lab does not give you detailed instructions
2. Use the model you developed in Part C that explores the relationship between ambient temperature and equipment performance.
Modify it to take into account any differences in the local ambient temperature.
You may use any combination of the technologies that you have been introduced to together with any other tools that you are already familiar with.
The final data set you generated in Part B may be useful.
3. Determine if any of the masts are exhibiting unusual behaviours.
Do you think it is significant?
❑ Yes ❑ No
4. Prepare a presentation that explains your methodology to a lay audience and presents your conclusions.
You may share this presentation with your fellow students and your instructor.