Group A Report

SCC-411

1 page

- Executive summary
- Overview **diagram** (Xavier)

2 page

- Relationship **diagram** (Iero)
- Design path (Short summary of Error file + All who has smth to say)

3-4 page

- Outliers + **graph** (Oleksa)
- 2.5 - one short paragraph (Loveen)
- 3.1 insights, conclusions (Jay)
- 3.4 final distribution specification (Oleksa)
- 3.5 two sentence conclusion + 3d **graph** (Iero)
- 4.2 **graph** + link + sentence about most busy time (Jay)
- 4.3 short Shiny description + link (Oleksa)

Note that the report itself will not be marked, but will be used as a reference document for marking the 3 remaining elements: group presentation, group demo, and team member interview. The marking schemes for these elements can be found here.
The page limit is 4 pages, with minimum font size 10pt.

Students and breakdown of carried out work

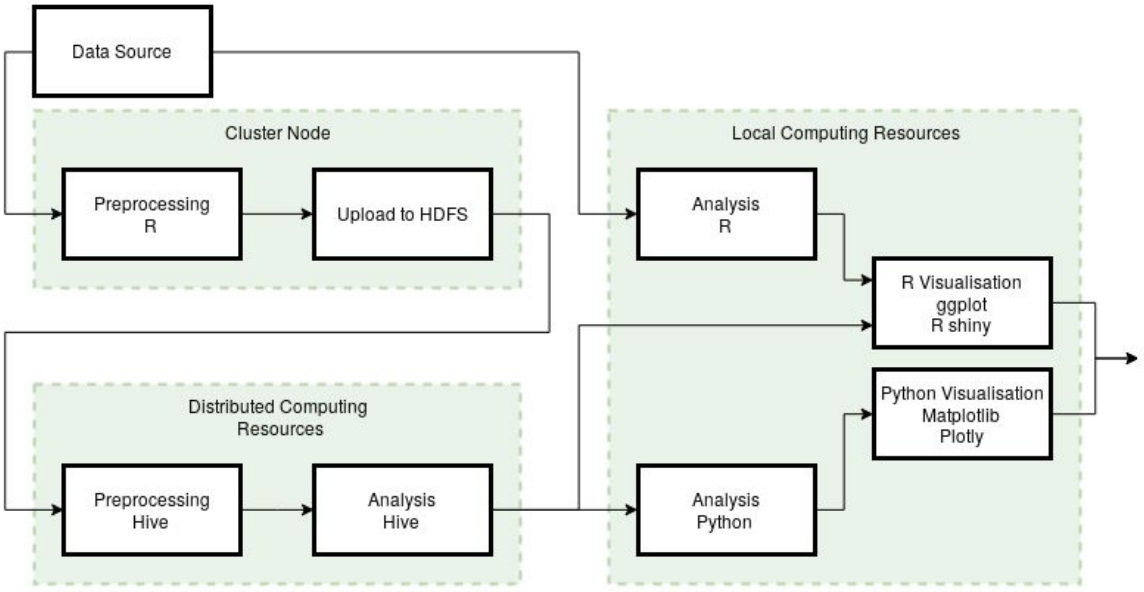
- Duquesne, Xavier - 35033737
- Dyall, Loveen - 32281487
- Mistry, Jayan - 32371684
- Stepaniuk, Oleksa - 32341885
- Tsantilas, Ierotheos - 32410961

Executive Summary (up to 400 words)

Running analysis of google servers failure on hadoop

Overview

<A high-level diagram of your system / pipeline> (see examples)



Schema

A ER diagram of your relational DB

Design Path

A list of issues you encountered, and how this affected your system design

Main outputs

2.1 (+4.1) Identify (and omit) outliers you come across in the data

Based on the analysis of the data we formed a hypothesis that 16 numerical variables from taks_events and task_usage tables roughly follow exponential distribution. If this is the case, on average 0.01% of observations should have value of more than Q3+7*IQR. This criteria was used to identify outliers (see example in the Figure XXX). To avoid losing the information we created two tables that identify outliers in each of the 16 variables in two tables. This allows to perform sensitivity analysis and exclude outliers when necessary.

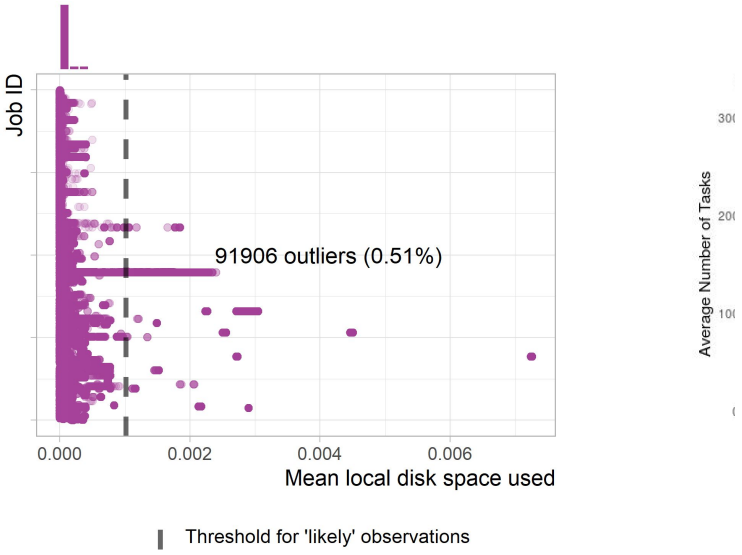


Figure XXX. Distribution of mean local space used

3.1 Coarse-grain analysis

3.1.1 Total number of unique users and scheduled jobs every minute

3.1.2 Average number of tasks per jobs

The most obvious observation is that there are more average tasks per jobs with restart than without, this is to be expected as the same task can be restarted multiple times, increasing the average amount of tasks that fail per jobs with restart. The average tasks per jobs is higher for lower priorities.

3.1.3 Total number of task failures

3.4 Create a probability distribution function of user job submission patterns

The average number of job submissions in the time interval *t depends* on the time trend and the number of submissions in the previous period. Model estimated with the Poisson

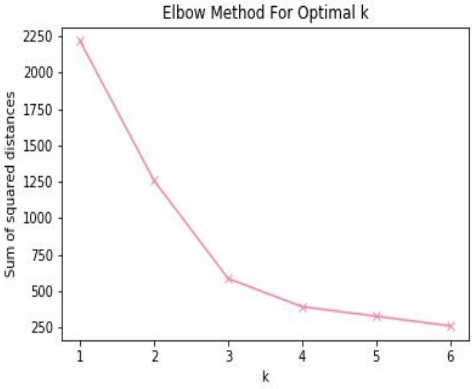
regression has highly significant coefficients (p-value<0.004), validity of the model is confirmed by the AIC criterion. Chi-square test shows that model is significantly different from the “ideal” case, but this is expected result given the small size of the sample. Final specification for the distribution is $Poisson(exp(2.67 + 0.75/log(t) + 0.01n_{t-1})$.

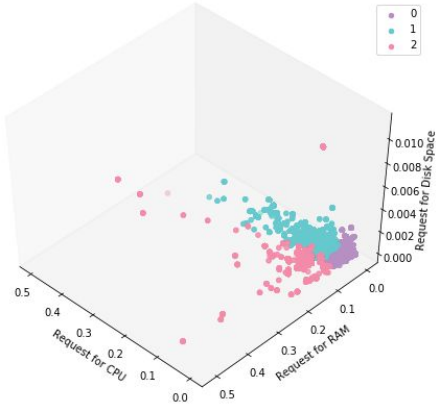
4.3 Online dashboard for comparing activity of the user and the average activity

Dashboard allows to select one of the users and provide summary statistics of his activity. In particular, it shows (1) the total number of jobs submissions in terms of priority in comparison to the median user; (2) dynamics of jobs submission in comparison to the average behavior; (3) comparison of the distribution for average resource usage across tasks between selected user and all other users; (4) table with description of jobs submitted by the user.
Link: https://ostepaniuk.shinyapps.io/users_dashboard/

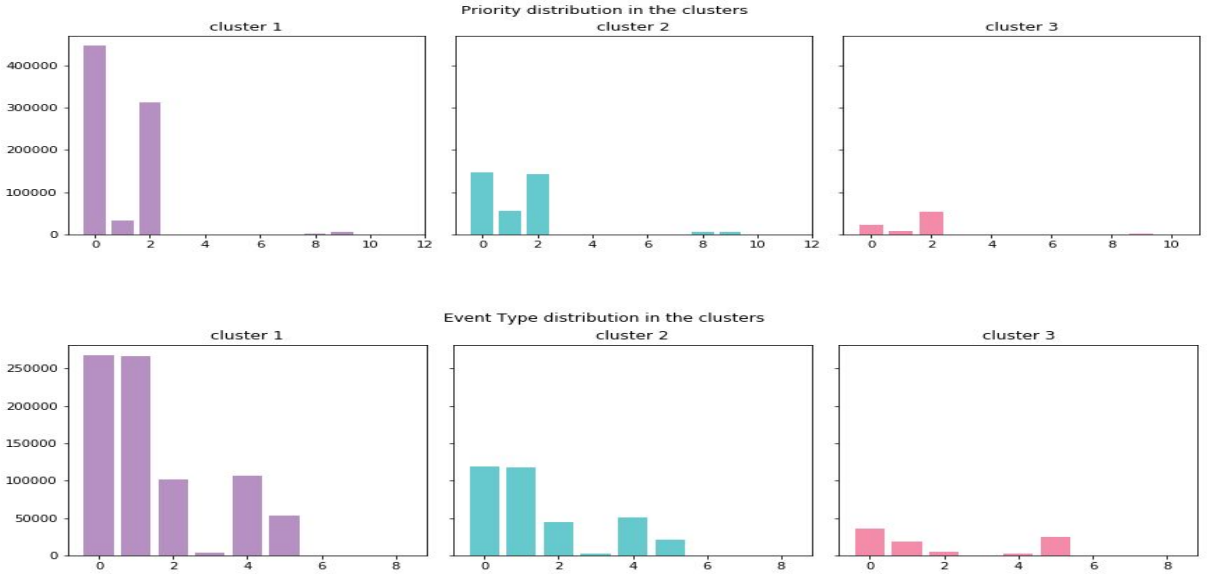
3.5 Creative: A K-Means Cluster Analysis for the task resource requests.

To examine if there is any relationship between the resources that requested for the tasks (limits) and some specific categorical variables like the task priority and the event type, we conducted K-means cluster analysis using the euclidean distance. We used the resources requested to feed our model (RAM, CPU, Disk Space). To estimate the optimal number of clusters we used the elbow method to see which number of clusters minimize the within-cluster distance (inertia).





From the clusters above we note that there only a few tasks that required more resources in terms of CPU, RAM and disk space and most of them belong to the third cluster. The next step is to analyse the behaviour of the priorities and event types in every cluster.



Regarding the distribution of task priorities we observe that in the first cluster most of the priorities are least important. These are in general tasks that require less resources. For the event types most of the tasks are rescheduled in the first cluster since these are tasks that have small priority. On the other hand, in the next two clusters the percentage of the tasks that were completed or killed increases significantly.

No relationship between cluster and structure of priority.
The tasks that request resources fail more often.