project

by Ismail Ahmed Nasser Al Jahdhami

Submission date: 08-Mar-2019 06:00PM (UTC+0000)

Submission ID: 101963535

File name: Project_Specification_and_Design_1.pdf (495.32K)

Word count: 4254

Character count: 21513

Department of Computer Science
2018/2019

One million Songs Dataset Project

Project Specification and Design

Big Data Group Project COMP530

Supervisor:

Dr. Michael Bane
Dr. Andrew Gargett (Hartree Centre)

By:

Al Jahdhami, Ismail	-	201281702
Arinto, Carlos	-	201386292
Yuefeng, Liang	-	201350099

1. Introduction

Dealing with very high volumes of data is a recurring problem and has become a concern to many organizations and entities around the world. It varies from textual data to audio and visual and [illegible] either be acquired through batch (already present) or streaming (real-time transmission) processes.

Notwithstanding, the musical field is a contributor to this emerging information factor, producing increasingly more audio files which are spread all over internet.

This project, titled "One Million Songs Dataset", intends to perform batch analysis on songs' metadata. This will allow the taking useful and specific outcomes such as "Which genre of music, due to its popularity, would be more likely to leverage the career of a newcomer artist?", creating a business model ready that could potentially be sold to these artists or any music-related party.

chk grammar

Some of the main challenges ahead will lie on cleaning the data, which will be done by a step-by-step process (first the data format is converted, followed by ignoring all unused fields, etc..), the batch analytics for which Apache Spark was the chosen framework, creating and populating the database done with Apache Hive. The visualization of data statistics will be possible via Jupyter Notebook interpreter

This project will be mainly developed in Python coding language and run on Hartree's HPC: Dawson, in order to increase its performance. Its final results will be made available in a web app to facilitate both the access and visualisation to those.

2. Aim

The aim of the project is to perform trend analysis on a set of data which describes and gives information about each of the one million songs (split between fields such as song name, beats per minute, etc..). to understand how the most listened music tastes have changed from 1922 to 2011 along with the creation of a business model, supported by finding trends on data which would, as an example, benefit companies planning on investing on the hottest artists.

As for the current state of the art, most projects who have worked on trend analysis of the "One Million Songs Dataset" have only performed a "MapReduce" operation over a single field[]. We will go a step further on this analysis, achieving conclusions, later on specified in the *Objectives,* only acquirable by relating the reduce outputs of multiple fields, for which Apache Spark will be the chosen tool.

3. Objectives

3.1 Essential Objectives

a) A clean dataset that can be organised by ascending or descending order regarding any given field.

b) Analysis of which artists / music genres are the most popular (hotness field) in each country.

c) Analysis of the genre and beats per minute of the most popular songs in the past few (5-10) years

3.2 Desirable Objectives

a) Develop a performance score based on popularity of the artists and their songs as well as the popularity of the genre.

4. Expected Outcomes (Deliverables)

4.1.Description of anticipated software

Essential:

- Store the data in Apache Hive data warehouse;
- Build a website to improve the accessibility and visualisation of the analysis performed on the data;
- Cleaning of the data by checking if it is on the right format, excluding any irrelevant information and dealing with missing values
- Use of Apache Spark framework to prepare the data to be analysed by converting it into key-value pairs
- Perform further analysis on the multiple parameters (in key-value format) which will be used to draw some of the "final stage" conclusions stated in the *Objectives*

- Use of a continuous integration environment such as GitHub to merge all the developer's work into a single workflow and provide version control along with team review. not an outcome

Desirable:

1- Use of Jupyter Notebooks to easily visualize statistics (such as graphs) derived from implemented algorithms or queries to the database;

2 – Perform unit tests that will assure that, whenever a posterior change is made, the code still works as intended not an outcome

Description of anticipated documentation:

Essential:

1 - Developer's guide: for the code portion of the project.

2 - User manual: directed at the visualisation of the conclusions taken regarding the mentioned *Objectives*.

Description of anticipated artefacts:

Essential:

1 - Diagrams that clearly summarize how the music tastes have changed over the years.

5. Resource requirements

Dataset:

The main resource of the project is one million songs' metadata (MSD), which is provided through a website from The Echo Nest. We will need to download the data and save them in the database. In addition, we also need to programmatically query the data. Wikipedia also provides songs metadata , so we could use some to replace the dirty data or fill in the missing features in MSD while in the cleaning data process. ~~Otherwise,~~ the data ~~is given to us~~ in HDF5 format, so we will ~~need to~~ use h5py package to manipulate data. The h5py package is a Pythonic interface to the HDF5 binary data format. Then, we will type python to transfer the dataset to csv format.

Hardware:

Due to the size of the dataset which is about 280 GB, it is required to process the data using high-performance computing resources that are available at Hartree, making use of multiple nodes to speed up the large-scale analysis of the process. For this project, Hartree provides access to Dawson which has the following Hardware Specification on the cluster:

- 3 Name Node servers with 16 cores and 128GB RAM
- 36 Data Nodes with 24 cores and 64GB RAM

The cluster will provide the required RAM, Hard disk storage and processing power to deal with the task.

Software:

In order to input code into the cluster a software framework called Hadoop is required. Hadoop is an open source framework for large-scale data process. We need to use the Hadoop Distributed File System (HDFS), which is a distributed file system of Hadoop. HDFS provides the distributed storage capabilities to process large data sets. Instead of using MapReduce, Apache Spark will be used due to its increased performance and usability ~~(Discuss the comparison in Research part)~~. Apache Hive will provide a querying language capable of querying on HDFS and map stored data files into a database table. In addition, many other software packages will need to be installed in case of using python, being the main ones PyHive and PySpark. Jupyter Notebook will ~~also need to~~ be used as the data visualisation tool, and Azure Databricks as a cloud services platform for test Apache Spark-based analytics. The final product will be available for consulting online, where the results will be gathered for the visualisation according to different planned objectives.

check grammar

6. Research

Music has had an important role in our culture throughout human history. The level of this importance is reflected in the music industry being worth billions of dollars. This industry relies heavily on deep understanding of what makes up a "popular" song, which is to say, a song whose value translates directly into revenue. Therefore, being able to analyse the trends in popular music would provide a significant capability to the industry; a key requirement for this capability to be realised is collecting enough music metadata to carry out the appropriate modelling. The Million Songs dataset more than adequately satisfies this requirement.

The Million Song Dataset (MSD) is an open resource for researchers, which we need to download metadata from. The reason why we choose MSD:

• large-scale, open source

- freely available
- easy to get started

In addition, there are also some metadata of songs in Wikipedia we could use for cleaning purpose.

6.1. Background literature

In 2012, there was a paper conducted by some researchers, which showed that some patterns and metrics describing the general primary musical features in modern western popular music, like pitch, timbre, and loudness, and it suggest three important trends in evolution of music: less variety in pitch progressions, more frequent timbres and louder volumes (Serrà, J. *et al.*, 2012).

6.2. Web programming documentation

For the website design, there are many tools to use. After conducting research to these tools, we may choose to perform by combining HTML, CSS or JavaScript. HTML is the standard markup language for creating web pages and web applications. We need to use it to describes the structure of a web page semantica check grammar ally included cues for the appearance of the document. HTML could also embed programs written in a scripting language such as JavaScript, which affects the actions and content of web pages. Besides, CSS is used to define the look and layout of content of the website.

6.3. Get familiarized on how to use the recommended big data tec[illegible]ogies

For this project, we need to look at performing analysis on large, disparate dataset. Research has been made into the big data stack to see what technologies will allow us to perform large scale data processing on these resources.

Hadoop + Spark

We need to look at options for processing the data. Firstly, Apache Hadoop is a cluster computing framework that enables distributed processing of large data on disk using a map-reduce model and allowing high-throughput access to the data using the Hadoop distributed file system (HDFS) for file management. In typical Hadoop task, MapReduce is used for the data processing, however, there are limitations to this as it persists the full dataset to the distributed file system after each MapReduce job, which means this is a very slow process.

Apache Spark also uses the HDFS for data management, but it introduced an in-memory caching and instead passes data directly without writing to persistent storage. Spark also speeds up data processing as it keep a JVM running on each node and it use DAG-based task scheduling mechanism, which better than MapReduce's iterative execution mechanism. Therefore, we prefer the Apache Spark rather than MapReduce which will be used to process the dataset. To use spark, team need to be trained on this as we are not very familiar on how to use it, so we will get online tutorials on how to use this technology.

Jupyter Notebook

As for data visualisation technique, there are various options to perform. We choose to use Jupyter Notebook as our data visualisation tool. The Jupyter notebook combines two components:

A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output. Notebook documents: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects.

After starting the Jupyter Notebook serves and input the dataset, we will type code on the interface to visualize the data.

6.4. Check on how to work with Apache Hive warehousing

We will install Apache Hive for querying large databases store in the HDFS. Apache Hive is a popular warehouse, which allows an SQL based language to convert SQL queries run on the Hadoop cluster. Because HiveQL is similar to SQL and Hive is faster, scalable and extension, we choose it for data query. Firstly, user PC collect data through Hive interface, and perform the execution job. Then, execution engine sends back the result to Hive interface. Because it is easy to operate and highly flexible, we will use it to query the data.

6.5. local Tools

For this project, we will use Pycharm software to type python code and download the Anaconda to use Jupyter Notebook locally. After we download the sample dataset, we will transfer the format to the csv and look into the data structure, which we could use . Also, we need MobaXterm to access Dawson via ssh.

7. Risk Assessment

For any success of a project it is mandatory to identify the risks that might the project encounter so it can be mitigated ~~as~~ earlier.

Hardware and Software Failure:

Any Failure of Hardware or Software due to parts or malicious attack can be catastrophic if there is no alternative way to run the project on or recover the data later on.

Schedule Risk:

There might be some delay to accomplish any schedule task, which might affect other tasks and delivery of the project

Technical Risk:

This could happen if there is some technical requirement that the team is not aware of that relate to programming or solving some issues in the project

Poor team dynamics:

Miscommunication or any problem with the team member, which may force them to leave the country which might affect the work related.

Changing of Scope:

Change of Hartree representative could bring new ideas and thought which might affect the project schedule and resources.

Moreover, we can measure the likelihood of risks, so we can mitigate their consequences that might affect the project performance. Below is the assessment of the risks that were identified:

	Risk Identified	Risk assessment Risk impact = Likelihood X Consequence	RAG Grading	Risk Mitigation
1	Hardware and software failure	3 X 4=12	Red (Critical)	Implement daily Backup and replicate of the code in the local pc and to google drive
2	Schedule risk	3 X 3 =9	Amber (Deserve some attention)	Adjust the time and try to fit it with another tasks schedule
3	Technical Risk	3 X 3 =9	Amber (Deserve Some attention)	Seek for help from Hartree or who has experience on that
4	Team poor Dynamics	2 X 3 =6	Amber (Deserve some attention)	Understand of team's behaviors and issues
5	Changing of scope	2 X 4=8	Amber (Deserve some attention)	Should be prepared to defend against excessive changes and additions once development has begun and be prepared to explain consequences.

8. Design:

8.1. System Overview

Considering that the full “One Million Song Dataset” contains around 280Gb of data, it would become too computationally expensive to perform any task on it. On the first stage of the project, personal computers will be the main tool to work on a much smaller subset (1.9Gb). When the complete dataset is used, Hartree Center’s HPC’s will provide the computational power necessary to operate the data within reasonable timestamps.

The data will be stored using Apache Hive, from where it will be queried and uploaded to ~~by~~ each personal computer and Hartree’s HPC:Dawson.

After the cleaning stage, it will be possible to visualize statistics about the data in any further point in the development by using Jupyter Notebook, ref required capable of displaying it in a graphical, more accessible and understandable way.

Lastly, the outcomes of this project will be mainly focused on creating a business model, useful for potential buyers, that could be handed in the form of an app. However, due to time limitations, this project's scope will display a sample of that model via a website where its user can access different angles of results (i.e, more oriented to festival sponsors, newcomer artists, etc...).

8.2. Data Sources:

The provided data contain the information of (track, song, release, and artist) which are in each file which basically have some redundancy, although the bulk of the data, relating to the audio analysis, is unique

In order to make analysis of the dataset that were provided by The Echo Nest we can analyses a subset of 10,000 songs (1%, 1.8 GB compressed) for a quick taste of the data to make testing locally or offline instead of analysing the whole dataset which is 280 GB of size.

The provided dataset is on HDF5 format which encompasses both metadata and audio analysis features. Each song is represented as a single file

The dataset will be loaded into Dawson by Hartree engineers where we can have access and start working on it. Moreover, the data that is provided is open source which is used for research purpose

8.3 Pre-process Design

8.3.1.Data cleaning

The whole dataset we have not check for, while we checked the sample dataset, and we found the data is not too dirty. There are missing values, extreme erroneous, incomplete data, redundant data and other forms of dirty data.

To deal with this we will have to perform several stages of data cleaning to deal with the different forms of data., including value imputation to deal with missing values, identifying and removing outlier values, removing the duplicate or irrelevant values.

Missing Values:

There are several types of data cleaning for missing values we will consider; we could ignore the missing values, use the attribute mean or use the mean from a sample of the whole data, by taking mean of points around the same time as the identified missing values. There are also other methods that involve dropping the missing values, but this the suboptimal because dropping data means losing information.

Duplicate or irrelevant Values:

For those duplicate or irrelevant data, we need to spend time to check the data file. Removing the unwanted values is the first step of cleaning process.

Missing categorical data:

The best way to handle this problem is to simply label them as “Missing”. We need to create a new class for these missing categorical values

Identify the Outlier

It may be hard to deal with identifying the outliers, because our data is all about features of songs, like artist name, song name. Very few is numeric data. We will find out some algorithms to solve this problem.

8.4. Data Storage and Access

The dataset will be stored at Dawson Cluster initially. Then the data will be loaded into HDFS to be analysed by python code.

After processing the data into Hadoop and got the output results the file will be loaded into Hive database and use HiveQL to be able to query the data that is stored into Hive database.

Loading the Dataset:

In order to access the dataset which resides in Dawson cluster Python programming language libraries are used. For instance will use the file "HDF5_getters.py", written by Thierry ertin-Mahieux at Columbia University. This file makes use of the python libraries numpy (Numerical Python) and tables (PyTables/Python Tables), which ~~aid to~~ deal with a hierarchical format such as HDF5. Also, we might use matplotlib/pylab for visualize the data in Jupiter. Moreover, data will be transformed to CSV files then be loaded into as HDFS files in Hadoop.

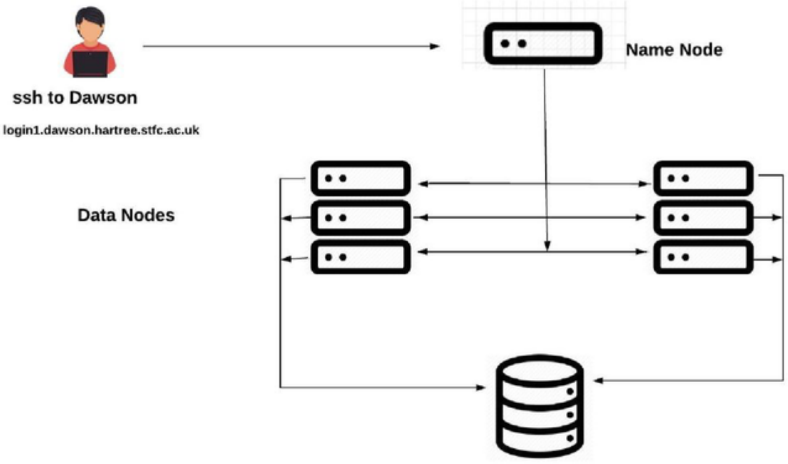


Figure 1: Process on how to access the data

ER Diagram:

A database ER diagram visualizes how entities within a database relate to each other and the attributes of each entity. For this project there is a need to create tables that helps to load the data into tables so it can be queried using HiveQL from Hive database. However, the structure of the tables might be modified during analysis process.

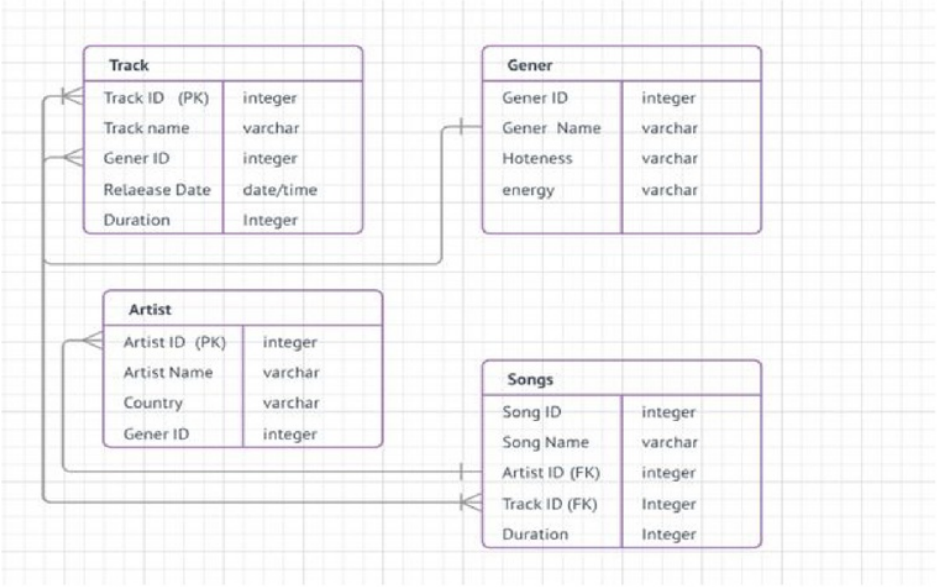


Figure 2 : ER diagram of database

8.6. Processing

The processing mode of the code will be made via batch analytics, using the MapReduce function to produce key-value pairs with respect to a given field.

First, a spark session is created and a name for the app can be provided

Then, we can proceed to the mapping of the keys with the method *map()*

Finally, through the *reduceByKey(<<operation>>)* method, the mapped values will be reduced into a single value by applying the operation specified in the parameter

Example for a WordCount.py program using PySpark:

```
spark = SparkSession.builder.appName("ExampleApp").getOrCreate()
lines = spark.read.text(sys.argv[1]).rdd.map(lambda r: r[0])
counts = lines.flatMap(lambda x: x.split(' '))\
                .map(lambda x: (x, 1)) \
                .reduceByKey(add)
```

This process is run every time we want to reduce with respect to a different key. In this project those keys will be the following (metadata fields): artist hotness, year of release, genre and song hotness.

8.7. Analysis and Performance Score

8.7.1 Analysis

8.7.1.1 Data cleanse

The first step to be able to perform the analysis is to clean the data.

This process will start developing and using a python algorithm which will convert the data from HDF5 format into CSV so that it can be uploaded into hive and allow the following cleaning methods. Next, all the unused parameter fields will be removed from the dataset. After that another algorithm will remove any symbols and punctuation from the existing textual fields.

Lastly, a final algorithm will remove any data point/instance which contains missing values. The dataset will then be ready to be uploaded and worked on

8.7.1.2 Most popular artists/genres in each country

The initial step for this analysis, will be to create group the artists which have identical values for the field “location”.

After this, the artists of each country will be ranked/ordered by popularity.

Now, in order to get the most popular genre, an algorithm will query and hold in memory all the songs created by the artists of each country, which translates separating having all the songs that “belong” to every country. Finally, the genres of these songs will be counted, culminating in having the most popular genre per country.

8.7.1.3 Genre and BPM of most popular songs in the past 10 years

Initially, an algorithm will pick all the songs between 2012 (most recent date of this dataset) and 2002.

Over that subset, the songs will be rated by popularity, by ordering them regarding the field "hotness".Then, all the songs that are on the "lower half" of this rank, will be discarded.

After this, a final algorithm will count the number of times each genre appears in this "most popular" subset while also computing the average of beats per minute of these songs.

The genre with the highest count value will then be picked, and the average of beats per minute of the top 50% most popular songs of the past 10 years will also be found.

8.7.1 - Performance Score

For the desirable objective, where a performance score which will merge scores from artist popularity, artist song's popularity respective genres popularity will be carefully developed, so that a more "in depth" analysis can be made without sacrificing the veracity of the final outcome.

chk grammar

8.8. System Evaluation

The system evaluation will assess whether or not the system can:

- ➢ Organise the artists by location and analyse the sum of their hotness with the average hotness of their songs. This will allow a sponsor to aim it's investment towards the festivals/concerts in which this artist participates.
- ➢ Gather all the songs that were produced from 2002 to 2012, check the average of beats per minute and most common genre. As this data is available to be consulted any interested musician can check which genre or rhythm would be best for him/her to pursue to attract more listeners.
- ➢ Rank the artists not just by hotness but also taking into consideration their songs and most used genre. For instance, the rank 1 artist should have a high score in "hotness",

have songs with a good “hotness” average, whose genre is amongst the most used/common. Advertising companies can then use these results to target the artists which are more likely to have the largest amount of visits on their websites, Youtube channels, etc… and augment the number of .the views the adds will get.

All information listed above will be displayed in the form of graphs and tables on the web application.

The system will firstly be evaluated by the personnel involved in the project and then shared with other parties such as musicians and workers in the fields of advertising and sponsoring who will provide feedback on the clarity and value/usefulness of the outcomes.

Their opinion shall determine if the aim of the project was fully fulfilled, partially fulfilled, or quite not yet satisfied, and should include constructive criticism both on the positive and negative aspects of the achieved results.

Use Case Diagram:

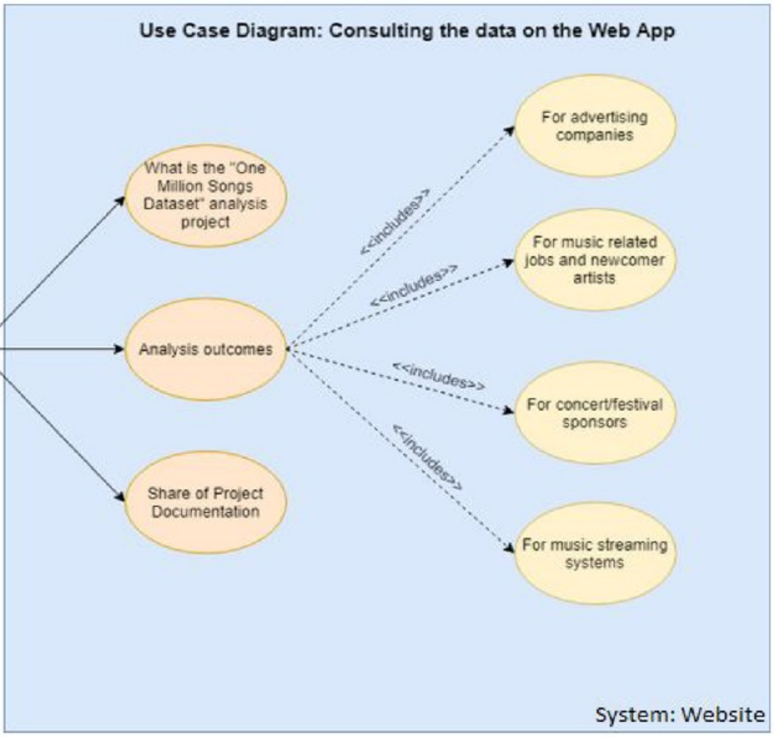
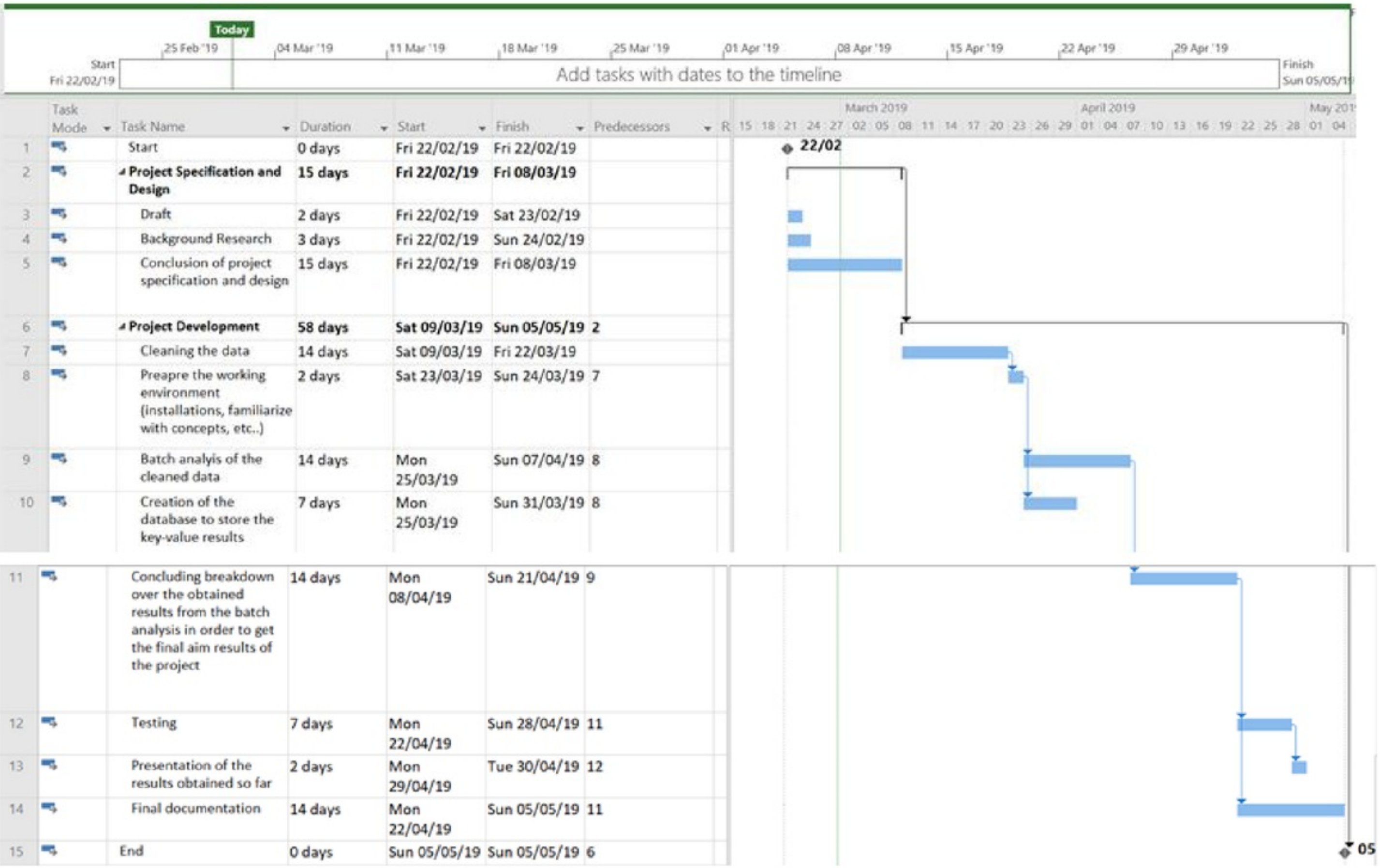

Use Case Diagram: Consulting the data on the Web App
User
What is the "One Million Songs Dataset" analysis project
Analysis outcomes
Share of Project Documentation
<<includes>>
<<includes>>
<<includes>>
<<includes>>
For advertising companies
For music related jobs and newcomer artists
For concert/festival sponsors
For music streaming systems
System: Website

Figure 3: Website design

Figure 3 illustrate on how the analysis can be used for other websites like for advertising , for concert/festival sponsors, music related jobs and new artists or for music streaming systems.

9. Plan



	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		Start	0 days	Fri 22/02/19	Fri 22/02/19	
2		**Project Specification and Design**	**15 days**	**Fri 22/02/19**	**Fri 08/03/19**	
3		Draft	2 days	Fri 22/02/19	Sat 23/02/19	
4		Background Research	3 days	Fri 22/02/19	Sun 24/02/19	
5		Conclusion of project specification and design	15 days	Fri 22/02/19	Fri 08/03/19	
6		**Project Development**	**58 days**	**Sat 09/03/19**	**Sun 05/05/19**	**2**
7		Cleaning the data	14 days	Sat 09/03/19	Fri 22/03/19	
8		Preapre the working environment (installations, familiarize with concepts, etc..)	2 days	Sat 23/03/19	Sun 24/03/19	7
9		Batch analyis of the cleaned data	14 days	Mon 25/03/19	Sun 07/04/19	8
10		Creation of the database to store the key-value results	7 days	Mon 25/03/19	Sun 31/03/19	8
11		Concluding breakdown over the obtained results from the batch analysis in order to get the final aim results of the project	14 days	Mon 08/04/19	Sun 21/04/19	9
12		Testing	7 days	Mon 22/04/19	Sun 28/04/19	11
13		Presentation of the results obtained so far	2 days	Mon 29/04/19	Tue 30/04/19	12
14		Final documentation	14 days	Mon 22/04/19	Sun 05/05/19	11
15		End	0 days	Sun 05/05/19	Sun 05/05/19	6

Milestones:

1. Project Specification and Design
2. Project Development
3. Testing

10. Bibliography

Serrà, J. et al. (2012) 'Measuring the evolution of contemporary western popular music'. doi: 10.1038/srep00521, (access on: 24 February 2019)

Kraus, N. (1992) 'Intuitive Toxicology: Expert and Lay Judgments of Chemical Risks', Risk Analysis, 12(2), (access on: 24 February 2019)

Millions Songs Dataset: [https://labrosa.ee.columbia.edu/millionsong/], (access on : 20 February 2019)

Lists of songs in Wikipedia: [https://en.wikipedia.org/wiki/Lists_of_songs/], (access on : 20 February 2019)

H5py for Python: [http://docs.h5py.org/en/stable/], (access on : 1 March 2019)

Hadoop: [http://hadoop.apache.org/docs/current/], (access on : 25 February 2019)

Tannir, K & Tannir, K n.d., 2014. "Optimizing hadoop for MapReduce. learn how to configure your hadoop cluster to run optimal MapReduce jobs", , [https://liverpool.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat00003a&AN=lvp.b4026123&site=eds-live&scope=site/], (accessed on : 1 March 2019)

Apache spark: [https://spark.apache.org/docs/latest], (access on : 23 February 2019)

Jupyter Notebook: [https://jupyter-notebook.readthedocs.io/en/stable/notebook.html], (access on 24 Feb 2019)

Apache Hive:[https://hive.apache.org/], (access on : 24 February 2019)

Html+CSS+JavaScript:[https://www.w3schools.com/], (access on : 24 February 2019)

Tomas, A ., "Song Analysis", [https://github.com/thomasSve/Million-Song-Dataset-Analysis/blob/master/report/song-popularity-prediction.pdf], (access on: 6 March 2019)

project

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

Comment 1

and data can either...

QM

chk grammar

check grammar

Comment 2

explain how you know this is a main challenge

Comment 3

Add paragraph explaining contents of this document

Comment 4

how?

Comment 5

need to break this sentence down so more easy to read

Comment 6

missing reference

PAGE 3

Comment 7

good to do more than previous people have done but what extra "wisdom" will you gain? compare theirs to yours

Comment 8

needs detail on how you would do this.

i could read the set of objectives as clear up some data and then just use the given "hotness" field to rank the results, but the project should be more than this

Comment 9

this reads a bit more like the steps to do something (which might be better in the Project Plan?)

Comment 10

This is rather muddled. You might be better to separate out software elements e.g.

Comment 11

"cleaning data" is design (verb) not a given outcome (noun). where you store data might also be considered a design choice.

your outputs will be

* cleaned-up data set: for use by others who want to write their own analysis routines

* a trained model based on this cleaned-up data set (for others who want to use it for ML)

* analysis software that offers a user a choice of 2 business model questions, uses the trained model and makes prediction for the selected business model question

then say software will implement x,z,y and be well commented etc etc

PAGE 4

QM

not an outcome

as per lectures, this is a Design choice (or possibly part of the "Plan" section

Comment 12

as written it's not an outcome. you could rewrite "A desirable outcome would be a (set of) Jupyter Notebooks that allow the user of our system to..."

QM

not an outcome

as per lectures, this is a Design choice (or possibly part of the "Plan" section

Text Comment. (MSD)

Comment 13

The data format of MSD is HDF5 format, and we have chosen to use the "h5py" package...

Strikethrough.

Strikethrough.

Comment 16

no comment

Text Comment. will

Strikethrough.

Comment 18

??

PAGE 5

Comment 19

REFERENCE

Comment 20

REFERENCES

Strikethrough.

QM

check grammar

always check your grammar makes sense

Comment 22

Need some references for some of these software apps

Strikethrough.

Comment 24

needs quantifiable data and appropriate references for this source of data

PAGE 6

Comment 25

needs another line or two of explanation

Comment 26

you need to state how this relates to your specific project

QM

check grammar

always check your grammar makes sense

Comment 27

I'd say 6.3 is more important that 6.2 so swap the order

Comment 28

some of the below should cite a source for more information, and where possible you should relate the technology to your specific project

PAGE 7

QM

chk grammar

check grammar

Comment 29

This is copied from the Jupyter Notebook web pages, which MUST therefore be referenced

QM

chk grammar

check grammar

Comment 30

How is the data to be visualised? Graphs or histograms or timeseries or...?

PAGE 8

QM

ref required

ref required

Comment 31

??

do you mean "to convert SQL queries to [be able to] run..."? what does it convert them to?

Comment 32

there's no "user PC" in your design

Comment 33

this is more DESIGN than it is RESEARCH

Strikethrough.

PAGE 9

PAGE 10

Comment 35

how will you identify which task/s risk bad scheduling?

Comment 36

mitigation #5 appears to be COPIED from

http://downloads.esri.com/support/ProjectCenter/Matrix_of_Common_Project_Risks.pdf

PAGE 11

Strikethrough.

QM

ref required

ref required

Comment 38

trend analysis has to be more than cleaning data and putting in a spreadsheet or table and then ranking by a given column - where is the "big data" here?

Comment 39

I thought you said it also had info such as beats per minute?

and your ER has "gener" ??

PAGE 12

Comment 40

how would you quantify?

Comment 41

you need to say specifically which you will do

Comment 42

why have you put it as 2nd step?

PAGE 13

Comment 43

say so in RESEARCH

QM

chk grammar

check grammar

QM

ref required

ref required

Strikethrough.

PAGE 14

Comment 45

I would remove this. It doesn't illustrate anything you are doing for your project. You would score a lot more by including pseudocode for your specific project

PAGE 15

Comment 46

where has this "hotness" come from?

Comment 47

repeat of above?

Comment 48

where does this field "location" get populated?

PAGE 16

Comment 49

WHY???

You need to justify why discard a proportion of data, and justify the size of that proportion

QM

chk grammar

check grammar

Comment 50

you could have some unit tests, eg to check all works with a subset of the data and then check all works with the full set of data

Comment 51

analysis of sum by average - not sure if that makes sense?

Comment 52

its

PAGE 17

Comment 53

this isn't part of the section on evaluation so should be in an earlier section

PAGE 18

PAGE 19

Comment 54

should indicate how team resources are assigned to different tasks

you have not discussed Easter break

you could include an activity diagram

PAGE 20

PAGE 21