Google/Apple Application Ranking Report

Extraction, Transformation, and Load (ETL) Specification Requirements

5/14/2019

Project Team:

- Satoko Yamaguchi
- Orathailux (Jerica) Saengara
- Humberto Rojas
- Trevor Laskey

Table of Contents

Introduction

The purpose of the Extraction, Transformation, and Load (ETL) Specification Document is to capture details that pertain specifically to ETL development to be used by the developer as an aid in ETL development.

1.1 Summary

Address in the section:

- Who is the business owner (The Client)?
  - Chris Shoe (Marketing Director) – Product Sponsor
  - Charity Faith Sotero (Sr. Marketing Manager) – Product User
  - Joshua Villamarzo (Sr. Marketing Manager) – Product User
- What is the business that the Client does?
  - Develops/manages social media marketing campaigns for mobile applications
  - Recommend new features/products based on industry trends
  - Performs quarterly reviews to senior management and trade organizations
- What business need, problem or objective will be addressed?
  - Identify the top selling applications for iOS and Android devices
  - Specifically interested in Google and Apple app store rankings
- Where do they want the data?
  - CSV file for internal reporting as available
  - Eventually file will populate our new "Marketing Dashboard" as potential KPI
- When do they need it by?
  - June 2019
- Were previous attempts made to integrate the data?
  - No attempts have been made to date

1.2 Scope of Initiative

Address in this section:

- The business process status
  - No changes to existing business processes are expected
- Expected business objective
  - Greater understanding of mobile application market
  - A measurement of consumer acceptance for new apps
  - Identify customer segments not being served by existing competitors
- Describe need for integration with any external system
  - Integration into "Marketing Dashboard" would help internal operations better identify those applications that have higher chance of achieving business/marketing objectives
  - Additionally, spot sudden trends and identify management when ranking thresholds are met.

1.3 Resources

Address in this section:

- Requesting Business Client
  - Project requirements and acceptance, UAT
- Project Manager or Business Analyst
  - Project team includes Digital Business Analyst
  - PMO access for internal project tracking
- Detailed specification of the operational source systems
  - TBD

1.4 Definitions, Acronyms and Abbreviations

Add items and terms that need to be defined in this section

Transformation Component Documentation

[This section outlines a more detailed description of the processes that are currently utilized and the proposed processes developed to achieve the objectives of this initiative.]

1.5 Data Import/Extract

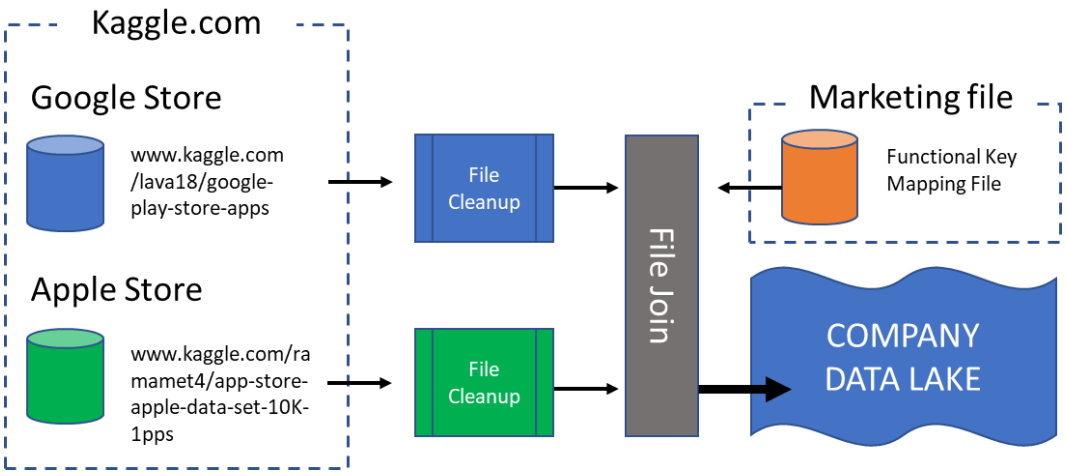
Address in this section:

- Database tables names
- File Exports names
- Transactions processes
- Special: parameter or control files

1.6 Data Import Method

Address in this section:

- Project ETL Process Flow Diagram



- Union in MySQL

```
CREATE DATABASE appstores_db;

USE appstores_db;

#SHOW CREATE TABLE googleplaystore;
#SHOW CREATE TABLE applestore;

#SELECT * FROM googleplaystore;
#SELECT * FROM applestore;

DROP TABLE IF EXISTS appstores_UNION;

CREATE TABLE appstores_UNION as
SELECT App as app_name, Category as category, Rating as rating, Reviews as rating_count, Price as price, 'google' as app_source FROM googleplay_cleaned
UNION
SELECT track_name as app_name, prime_genre as category, user_rating as rating, rating_count_tot as rating_count, price, 'apple' as app_source FROM applestore_cleaned
;

SELECT * FROM appstores_UNION;
```

- Functional Key Mapping (used to create the marketing file for file join process)

Apple Store	Google					
Book	BOOKS_AND_REFERENCE					
Business	BUSINESS	ART_AND_DESIGN	EVENTS			
Catalogs --> Shopping						
Education	EDUCATION					
Entertainment	ENTERTAINMENT	COMICS				
Finance	FINANCE					
Food & Drink	FOOD_AND_DRINK					
Games	GAME	FAMILY				
Health & Fitness	HEALTH_AND_FITNESS					
Lifestyle	LIFESTYLE	BEAUTY	DATING	HOUSE_AND_HOME	PERSONALIZATION	PARENTING
Medical	MEDICAL					
Music --> Entertainment						
Navigation	MAPS_AND_NAVIGATION					
News	NEWS_AND_MAGAZINES					
Photo & Video	PHOTOGRAPHY	VIDEO_PLAYERS				
Productivity	PRODUCTIVITY					
Reference --> Book						
Shopping	SHOPPING					
Social Networking	SOCIAL	COMMUNICATION				
Sports	SPORTS					
Travel	TRAVEL_AND_LOCAL	AUTO_AND_VEHICLES				
Utilities	LIBRARIES_AND_DEMO					
Weather	WEATHER					

1.7 Data Acquisition

Address in this section:

- What data is needed
  - Kaggle - Apple Store
  - Kaggle - Google Play Store
- Will technical data specifications be provided?
  - Comma Separated (CSV)
- How will the data sets be identified and selected?
  - Based on availability
  - Cost
- How often are source systems updated?
  - N/A
- How much data can be expected?
  - < 1 gb
- How often will the data be received?
- What state is the source system?
  - Production?
  - Testing – when will the design be finalized?
- When can a production extract be ready for use by the developer?
- Are there any specific bottlenecks to getting the data?

1.8 Data Transform Specification

Address in this section:

- Has any technical analysis been previously performed?
  - No
- Do design specification or data models exist?
  - No
- What kind of history or trending information is needed?
  - Client requests one year look-back
- Are there any known fields that require a derived calculation?
  - Only the FK mapping file

1.9 Data Integrity

Address in this section:

- How reliable is the source system data (e.g., empty fields, dates stored as text, invalid code values, text fields with odd or control characters, etc.)?
  - The raw data files come from Kaggle service
  - These files are maintained by user community
  - May need to find alternative data set in future
- Are the source data sets used by more than one data owner?
  - No
- How often does the source system change?
  - Quarterly (or as required)
- How will changes be communicated to the users and development team?
  - TBD

1.10 Data Refresh Frequency

Address in this section:

- How often should the Data Warehouse process refresh?
  - Quarterly (as available)

1.11 Data Security

Address in this section:

- Do monitoring requirements need to be satisfied?
    - Yes, specifically when data categories change as they are our only means to join tables using functional key (category)
- Are there any federally mandated HIPAA considerations?
    - No
- Is there need to build in additional privacy?
    - Privacy issues are addressed in dashboard requirements definition document

1.12 Data Target/Result

Address in this section:

- How long will be data retention?
- How often can the data be purged?
- What interface will the Client/Users access the data?
    - Reports
    - Dashbords
- Does the project data have an expiration date?

Data Quality

1.13 Performance Measures and Standards

Address in this section:

- What measure will determine the success of this project?
- Do specific (Federal or otherwise) compliance measures exist?
- Do other specific outcomes exist?
    - Expected Reports
    - Totals and expected counts
    - Technical comparison (e.g., using SQL, etc.)

1.14 Resources

Address in this section:

- Who will be performing the user acceptance testing?
- Do they have a timeline or window for testing?
- When will they be doing it?