**Datasets (mostly from Kaggle) and Example Analytics**

**COVID-19 Cases, Deaths Global Data**

* <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>
* Size: 17 MB [all the data from Jan 22 to today, updated daily]

**Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE**

* <https://github.com/CSSEGISandData/COVID-19>

**Data Science for COVID-19 (DS4C)**

DS4C: Data Science for COVID-19 in South Korea

* <https://github.com/ThisIsIsaac/Data-Science-for-COVID-19>
* **Example analytics (**they are not based on any specific dataset, but just for illustration purpose**)**
  + What is the ranking/distribution of #active cases/first case date/#death/fatality rate by of countries/states/cities?
  + Which countries/states have flattened the curve (e.g., measured by changes in growth rates)?
  + How long does it take for a patient to recover? How is this changed over time?
  + Are there any correlations between infection and age groups/gender?

**Uber Pickups in New York City**

Trip data for over 20 million Uber (and other for-hire vehicles) trips in NYC

* <https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city>
* size: 835MB
* files: 19
* **Example analytics (**they are not based on any specific dataset, but just for illustration purpose**)**
* What are the peak hours for Uber pickups during the day?
* What are the top-3 popular streets that get the most pickup requests?
* How do popular streets change overtime during the day?

/\* this may help drivers find the right place to go at the right time \*/

**NY Parking Violations Issued**

Explore Open Data from New York City

* <https://www.kaggle.com/new-york-city/ny-parking-violations-issued>
* Size: 5GB (one-year data is around 1GB)

**NFL Play Statistics dataset (2004 to present)**

NFL play-by-play dataset with participation information, 2004-present.

* <https://www.kaggle.com/toddsteussie/nfl-play-statistics-dataset-2004-to-present>
* size: 476MB
* columns: many

**FIFA 19 complete player dataset**

18k+ FIFA 19 players, ~90 attributes extracted from the latest FIFA database

* size: 9MB
* columns: 89

**NBA games data**

Dataset with all NBA games from 2004 season to Feb 2020

* <https://www.kaggle.com/nathanlauga/nba-games>
* size: 91MB
* files: 5

**Dota 2 Matches**

Explore player behaviour and predict match outcomes.

* <https://www.kaggle.com/devinanzelmo/dota-2-matches>
* files: 18
* size: 1GB

**Food.com Recipes and Interactions**

Crawled data from Food.com (GeniusKitchen) online recipe aggregator

* <https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions>
* size: 851MB
* columns: (6, 6, 6, 8, 6, 5, 12)

**US Accidents (3.0 million records)**

A Countrywide Traffic Accident Dataset (2016 - 2019)

* <https://www.kaggle.com/sobhanmoosavi/us-accidents>
* size: 1GB
* columns: 49

**Madrid Airbnb Data**

Information about Airbnb Listings in Madrid, Spain

* <https://www.kaggle.com/rusiano/madrid-airbnb-data>
* size: 613MB
* files: 7

**Air Quality Annual Summary**

A summary of air quality from 1987 to 2017

* <https://www.kaggle.com/epa/air-quality>
* columns: 55
* size: 948MB

**380,000+ lyrics from MetroLyrics**

Lyrics, Artist, Genre, Year

* <https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics>
* columns: 6
* size: 310MB

**World of Warcraft Avatar History**

Track the players of this popular online game

* <https://www.kaggle.com/mylesoneill/warcraft-avatar-history#zones.csv>
* size: 614MB
* files: 4

**JCPenney products**

20,000 product listings from JCPenney

* <https://www.kaggle.com/PromptCloudHQ/all-jc-penny-products>
* columns: 14
* size: 23MB

**Trending YouTube Video Statistics**

Daily statistics for trending YouTube videos

* <https://www.kaggle.com/datasnaek/youtube-new>
* size: 514MB
* columns: 16

**TMDB 5000 Movie Dataset**

Metadata on ~5,000 movies from TMDb

* <https://www.kaggle.com/tmdb/tmdb-movie-metadata>
* size: 44MB
* columns: 20+4

**Kickstarter Projects**

More than 300,000 Kickstarter projects

* <https://www.kaggle.com/kemical/kickstarter-projects>
* size: 100MB
* columns: 16+15

**Pick your own dataset**: [https://www.kaggle.com](https://www.kaggle.com/kemical/kickstarter-projects) Please consider these when you select a dataset

* sufficiently large for performance testing (ideally larger than 100MB)
* rich in structure (multiple CSV files and multiple columns)

The data store app will be required to provide these features by the end of the quarter:

(1) support for the basic search for fields of interests, update/insert/delete records in the data store;

(2) support for backing up the latest data store and importing backup data;

(3) support for 6-8 data analytics (like the ones mentioned in the Uber and COVID-19 examples), depending on the complexity of each analytics;

(4) support for some incremental analytics, where re-computing the analytics can be done more quickly after an update than running from scratch