Class: As said in class, the focus of your projects is on how to model, store, and process a large scale of variety of data. The expectation is that you will explore additional data types (e.g., images, time-series, geospatial, streaming data, etc.), additional NoSQL databases (other than we see in class) that are suitable for managing these data, and explore the parallel processing capability of these databases and/or the utilization of Spark. The purpose of this document is to help you brainstorm and form project groups. Please find real-world use cases for your project (see examples below). More detailed requirements will be provided later..

This document is meant for you to contribute and share with others. I will start with a table listing interesting data types, web sites or papers, and use cases. Please contribute to this table and indicate your names after the resources you contribute. Then in the second table, I will ask you all to propose & collaborate on possible project topics.

|  |  |
| --- | --- |
| **Data type** | **Resources** |
| Image/audio/video | * [PostgreSQL Application in Image Search and Video and Image Deduplication](https://www.alibabacloud.com/blog/postgresql-application-in-image-search-and-video-and-image-deduplication_595794) * *Pillow:*[Python Imaging Library](https://pillow.readthedocs.io/en/stable/) * B[asic Image Data Analysis Using Python](https://www.kdnuggets.com/2018/09/image-data-analysis-python-p3.html) * [10 Python image manipulation tools](https://opensource.com/article/19/3/python-image-manipulation-tools) * [An introduction to audio processing and machine learning using Python](https://opensource.com/article/19/9/audio-processing-machine-learning-python) * [Audio processing in Python (short tutorial)](https://publish.illinois.edu/augmentedlistening/tutorials/music-processing/tutorial-1-introduction-to-audio-processing-in-python/) * Please contribute:   + ... |
| Time-series | * [Advanced Time Series with Cassandra](https://www.datastax.com/blog/advanced-time-series-cassandra) * [Eye or the Tiger: Benchmarking Cassandra vs. TimescaleDB for time-series data](https://blog.timescale.com/blog/time-series-data-cassandra-vs-timescaledb-postgresql-7c2cc50a89ce/) * [How to efficiently store and query time-series data](https://medium.com/@neslinesli93/how-to-efficiently-store-and-query-time-series-data-90313ff0ec20) * [Time-series data: Why (and how) to use a relational database instead of NoSQL](https://blog.timescale.com/blog/time-series-data-why-and-how-to-use-a-relational-database-instead-of-nosql-d0cd6975e87c/) * [Cassandra Time Series Data Modeling For Massive Scale](https://thelastpickle.com/blog/2017/08/02/time-series-data-modeling-massive-scale.html) * [Time Series Analysis with Spark](https://databricks.com/session/time-series-analysis-with-spark) * [Time Series Analysis in Python: An Introduction](https://towardsdatascience.com/time-series-analysis-in-python-an-introduction-70d5a5b1d52a) * [An End-to-End Project on Time Series Analysis and Forecasting with Python](https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b) * [Time Series Analysis in Python – A Comprehensive Guide with Examples](https://www.machinelearningplus.com/time-series/time-series-analysis-python/) * [Analyzing Financial Time Series Using BigQuery and Datalab](https://cloud.google.com/solutions/time-series/analyzing-financial-time-series-using-bigquery-and-cloud-datalab) (but note Google cloud & bigquery usage may not be free). * Please contribute:   + ... |
| Geospatial | * [PostGIS is a spatial database extender for PostgreSQL object-relational database](https://postgis.net/) * [Processing Geospatial Data at Scale With Databrick](https://databricks.com/blog/2019/12/05/processing-geospatial-data-at-scale-with-databricks.html)s (Spark) * [Geopandas](https://geopandas.org/): working with geospatial data in Python * [Geospatial options in Apache Spark](https://databricks.com/session_na20/geospatial-options-in-apache-spark) * [Essential geospatial Python libraries](https://chrieke.medium.com/essential-geospatial-python-libraries-5d82fcc38731) * [Geospatial Analysis with Python and R](https://kodu.ut.ee/~kmoch/geopython2018/) * Please contribute:   + ... |
| Streaming data | * [Using PostgreSQL for Real-Time IoT Stream Processing Applications](https://medium.com/dataseries/using-postgresql-for-real-time-iot-stream-processing-applications-965741c57315) * [Discretized Streams: Fault-Tolerant Streaming Computation at Scale](https://cs.stanford.edu/~matei/papers/2013/sosp_spark_streaming.pdf) (Spark) * Spark Streaming: What Is It and Who’s Using It? (e.g., continuous ETL pipeline at Uber and Pinterest; data may come from [Apache Kafka](https://kafka.apache.org/)). * [Apache Kafka: open-source distributed event streaming platform](https://kafka.apache.org/) * Please contribute:   + ... |
| Graph data | * [Graphx: A resilient distributed graph system on spark](https://dl.acm.org/doi/pdf/10.1145/2484425.2484427) * [Big Data Processing Using Apache Spark - Part 6: Graph Data Analytics with Spark GraphX](https://www.infoq.com/articles/apache-spark-graphx/) * [networkx: Network analysis in Python](https://networkx.org/) * [Amazon Neptune: fast reliable graph database for the cloud](https://aws.amazon.com/neptune/) ([YouTube](https://www.youtube.com/watch?v=6dHiJSMdbjc)) * [Graph analytics with graphx](http://ampcamp.berkeley.edu/big-data-mini-course/graph-analytics-with-graphx.html) (Berkeley data science bootcamp) * [Spark GraphX Tutorial – Graph Analytics In Apache Spark](https://www.edureka.co/blog/spark-graphx/) (flight data analysis) * Please contribute:   + ... |
| Key-value | * [Top Redis Use Cases by Core Data Structure Types](https://scalegrid.io/blog/top-redis-use-cases-by-core-data-structure-types/) * Please contribute:   + ... |
| Wide-column data or columnar data | * [Cassandra Data Modeling Best Practices](https://tech.ebayinc.com/engineering/cassandra-data-modeling-best-practices-part-1/) (eBay)  [MariaDB ColumnStore](https://mariadb.com/kb/en/mariadb-columnstore/)[ColumnStore Architecture & Use-case](https://mariadb.com/de/resources/blog/columnstore-architecture-use-case/)  * [Hive - A Warehousing Solution Over a Map-Reduce Framework](http://www.vldb.org/pvldb/vol2/vldb09-938.pdf) * Please contribute:   + ... |
| Use cases | * [How Twitter Uses Redis To Scale - 105TB RAM, 39MM QPS, 10,000+ Instances](http://highscalability.com/blog/2014/9/8/how-twitter-uses-redis-to-scale-105tb-ram-39mm-qps-10000-ins.html) * [Real world use cases for DynamoDB](https://d1.awsstatic.com/events/reinvent/2019/REPEAT_1_Real-world_use_cases_for_Amazon_DynamoDB_DAT305-R1.pdf) * Please contribute:   + ... |

Here, you can suggest your project idea, data types and databases used, and how the project addresses the scalability problem (e.g., leverage existing capabilities of NoSQL database used, or use Spark for processing, and store processed data back to database, etc.), and interested people. Please be creative and link to your background/discipline, and avoid common data sets and projects that are similar to that in past semesters (e.g., searching IMDB/airbnb/yelp/NBA etc data).

|  |  |  |
| --- | --- | --- |
| **Project idea** | **Data types, databases, scale** | **Interested people** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |