Shubhangi Singhal (ss100)

IS 497 Database Admin & Scaling Final Project Proposal

MongoDB using Amazon Web Services

MongoDB is an open-source document database and leading NoSQL database. It is a cross-platform, document-oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

Technologies –

1. Database Engine - MongoDB Engine
2. Cloud Platform – Amazon Web Services (EC2)
3. MongoDB Compass (Visual tool/ Client for MongoDB)

Dataset –

1. Name – food-inspections.csv (222.08 MB)
2. Link – https://www.kaggle.com/chicago/chicago-food-inspections/activity
3. Description – This data is derived from inspections of restaurants and other food establishments in Chicago from January 1, 2010 to the present. Inspections are performed by staff from the Chicago Department of Public Health's Food Protection Program using a standardized procedure. The results of the inspection are inputted into a database, then reviewed and approved by a State of Illinois Licensed Environmental Health Practitioner (LEHP).
4. Metadata – There are 196825 rows and 22 columns. Column names – Inspection ID, DBA Name, AKA Name, License#, Facility Type, Risk, Address, City, State, Zip, Inspection Date, Inspection Type, Results, Violations Longitude, Latitude, Location, Historical Wards 2003-2015, Zip codes, Community Areas, Census and Wards.

Project Outcomes –

1. Setting up of MongoDB engine using Amazon Web Services and MongoDB Compass.
2. Performing rudimentary configurational changes basis, the requirements of the targeted database and project goal.
3. Implementing security policies for data transfer.

4. Loading of dataset and querying records.
5. Running backup and recovery procedures.

As per the project goal which is to utilize a database engine type I have not used yet this semester, I decided to experiment with MongoDB as my database as it is one of the most popular NoSQL databases which exist today. MongoDB goes well with the chosen dataset because we have data about 222 MB and this database engine works well with large amount of data. Moreover, it provides Document Oriented Storage where the data is stored in the form of JSON style documents. Since this dataset gets updated every Friday, we required a tool which is easy to scale and provides fast in place updates. Other than this MongoDB offers to have no complex join mechanisms and supports dynamic queries on documents using a document-based query language that is nearly as powerful as SQL. Finally, it uses internal memory for storing the working set, enabling faster access of data.

References –

1. https://www.mongodb.com/
2. https://aws.amazon.com/quickstart/architecture/mongodb/
3. https://www.tutorialspoint.com/mongodb/mongodb_overview.htm
4. https://www.kaggle.com/chicago/chicago-food-inspections/activity
5. https://data.cityofchicago.org/api/assets/BAD5301B-681A-4202-9D25-51B2CAE672FF