ST 590 Project 2 – Part 1

Claudia Donahue, Collin Knezevich

Uber is one of the fastest growing companies in recent history. According to Statista, Uber's total global revenue increased from $100 Million in 2013 to $6.5 Billion in 2016 – a 6400% increase over just 4 years. Today, Uber generates nearly $17.5 Billion in revenue and stores approximately 100 PB of data. As Uber began to grow rapidly, they recognized the need to efficiently store and access a large amount of data. Their initial solution was to use a data warehouse software called Vertica. While this platform was successful in allowing Uber's many data analysts and engineers to easily access and query data, it presented a number of data problems. First, data reliability was not very good under this platform. The proper schema for data was not formally defined; this combined with ingesting data through ETL jobs often resulted in duplicate data. Additionally, scalability needed to be improved, and this was a major problem given the large amount of data Uber takes in.

To address these issues, Uber switched to an Apache Hadoop-based data lake platform, which is still the basis of their big data operations to this day. The initial switch to Hadoop improved upon these issues with scalability and reliability; however further improvements would still be necessary in the future. Regardless, Uber was in a much better place regarding these issues. A major reason for these improvements is the usage of Apache Parquet, a data file format available in Hadoop that is based on a column-oriented format. Using this allowed Uber to improve its storage via better compression, resulting in better scalability.

Additionally, the Hadoop platform allowed users to use a number of tools to help them easily query and access data through a single user interface. Firstly, Presto allowed users to make

interactive queries. Next, Apache Spark assisted users in accessing raw data. This tool is very flexible, allowing users to access data via SQL or via other non-SQL methods. Finally, users could utilize Apache Hive for very large queries. Having all of these technologies available is very important so that individual users will have something to use that is tailored to what they need to query or do.

There were still some limitations with Uber's big data platform, and they sought to address these by adding on to their Hadoop-based platform. A major problem Uber faced was with data latency. Uber needed to make analytical decisions in real-time, but new and updated data only became available every 24 hours. As a result, ETL jobs were quite slow, as they needed to access the entire table every time. To remedy some of these issues, Uber added Hudi to their big data pipeline. Hudi added support for update and delete operations, vastly improving Uber's data latency to under an hour. The ability to update data much more quickly improved Uber's ability to make real-time decisions. Additionally, queries were now much more efficient, and had additional capabilities as well. Users could still query all records in a table at a point in time, but the queries ran much faster. In addition, users could choose to return only new or updated records in a table. Uber added a few supplemental technologies alongside Hudi. Firstly, they added Apache Kafka, which improved upon the storage of changelogs and their associated metadata. Uber also added Marmaray to their platform, which improved the efficiency of data ingestion by helping to avoid ingesting duplicate data. Data transformation is no longer done upon ingestion, but is done through Hadoop by users.

Uber relies on its data to answer several questions. Generally, the focus is how to lower costs, increase revenue, and improve customer experience. The immense amount of data the company possesses is useful in forecasting how much demand to expect at all times, how to

convert more job applicants into Uber drivers, and in detecting and preventing fraudulent payments on its platform, for example.

A critical goal the company has is to accurately forecast surges in Uber requests. During busier times, Uber increases its prices for riders and pays a bonus to its drivers. Predictable surges during weekends and rush hour are not as complicated to forecast as rare, extreme events due to events like New Year's Eve, weather catastrophes, or big concerts. Uber's engineers have developed a custom time series forecasting model to train a neural network to accurately forecast demand around such events. Uber acquires the necessary data for this analysis by tracking data from its drivers, both when they are transporting a passenger and when they are not. This allows them to analyze traffic patterns, an important component of surge pricing, among other things. In fact, they are conducting research on autonomous cars using the data acquired from its drivers. Additionally, Uber does research on the quality and availability of public transportation in different cities. These insights help them determine which cities to allocate more or less of their resources towards.

Another question whose answer Uber has pursued with its data is how to get more drivers to complete their sign-up process. The company began tracking the steps required to apply and be hired as a driver, and then analyzed the data to better understand the process, where they were losing applicants, and how to achieve a higher conversion rate. The tool the company's visualization team ended up developing is known as Maze. It is a visualization of a sunburst, where full and partial rings are made up of events during the hiring process. An example is uploading one's driver's license. Uber has determined that 20 percent of applicants' first attempt to upload a photo of their license were not successful. Only 57 percent of this group actually tried to upload the license a second time. Uber uses Maze to find root causes of problems.

Uber processes thousands of financial transactions every second, so another important question to answer is how to identify fraudulent payments. To correctly detect fraud and avoid false accusations, the company has noted the importance of keeping humans involved in the process, while relying on artificial intelligence to initiate review by a human analyst. The detection and mitigation Uber uses for financial fraud is called RADAR. When RADAR finds a potential attack, the company prioritizes it by the amount of the potential financial loss to the company, so they can focus on the most severe.

A large component of Uber's data analytics goals is to improve customer experience and to provide customers with information. For instance, Uber uses predictive models in order to estimate when the customer's driver will arrive, and when they will reach their destination. Not only does Uber seek to improve the experience of its customers – they also seek to improve the experience of their drivers. This data will also be relayed to the drivers to help them decide whether or not to accept a ride request. Additionally, Uber provides its drivers with heat maps in order to help them understand where demand is highest at a given point in time. Behind the scenes, Uber's matching algorithms help direct a ride request from a customer to the optimal driver, considering a number of variables. Correct identification of a driver that is likely to accept a request benefits both parties, ensuring that the service is performed as quickly as possible.

Many of these insights can also be used by Uber's other services such as Uber Eats. However, the food delivery model presents some new challenges. In order to minimize time waste, the delivery driver should arrive at the restaurant the moment the food is ready to be delivered. Uber uses machine learning models to try to predict the time it takes for a certain restaurant to prepare their food (also considering the specific dish being prepared). This also

means that they need to match drivers optimally, by sending them to restaurants a certain distance away so that they arrive at the right time (the time at which Uber predicts the order will be ready). Getting real-time data about restaurants is difficult, since they are not directly partnered with Uber. Thus, Uber uses historical data based on the past history of restaurants, as well as real-time data from its drivers that can roughly predict the current state of restaurants in an area.

References

"How Uber Uses Data Science To Reinvent Transportation?". *Projectpro*, 2022, https://www.projectpro.io/article/how-uber-uses-data-science-to-reinvent-transportation/290.

"Predicting Time To Cook, Arrive, And Deliver At Uber Eats". *Infoq*, 2022, https://www.infoq.com/articles/uber-eats-time-predictions/.

"Uber: Net Revenue Worldwide 2021 | Statista". *Statista*, 2022, https://www.statista.com/statistics/550635/uber-global-net-revenue/.

Laptev, Nikolay et al. "Engineering Extreme Event Forecasting At Uber With Recurrent Neural Networks". *Uber Engineering Blog*, 2022, https://eng.uber.com/neural-networks/.

Luo, Yujia, and Jerome Cukier. "Maximizing Process Performance With Maze, Uber's Funnel Visualization Platform". *Uber Engineering Blog*, 2022, https://eng.uber.com/maze/.

Zelvenskiy, Sergey et al. "Project RADAR: Intelligent Early Fraud Detection System With Humans In The Loop". *Uber Engineering Blog*, 2022, https://eng.uber.com/project-radar-intelligent-early-fraud-detection/.