

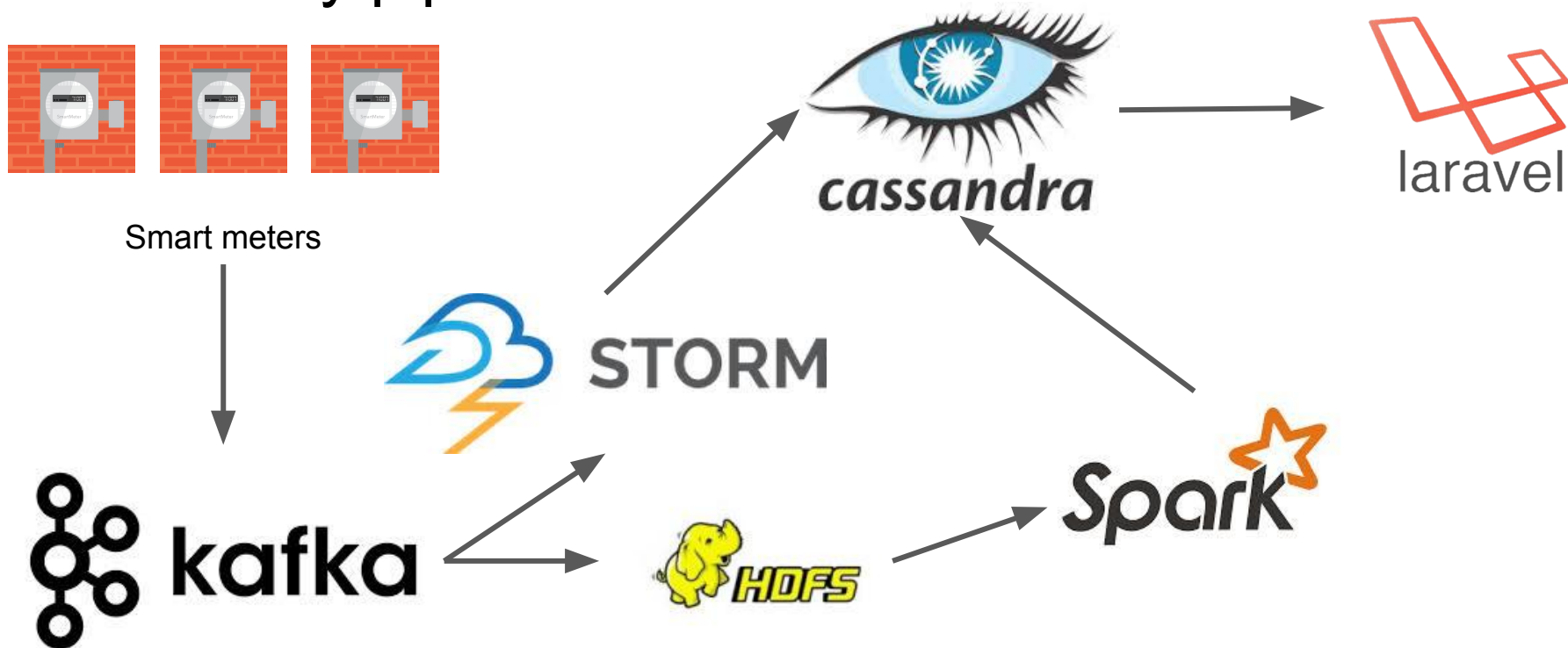
The project you probably already heard

Insight Data Engineering, New York

Motivation

- Problem:
 - Current utility systems are outdated and managing these systems is largely manual.
 - Utility systems are increasingly getting connected and existing systems are used for monitoring at energy companies.
- Business Use Cases:
 - Real-time monitoring and managing current usage
 - City wide analytics for consumption
 - Reactive triggers to notify anomalies

Preliminary pipeline



Queries on data

- Some queries:
 - What is my usage over time?
 - Last month
 - Last week
 - Last hour
 - Is my usage more than the usage of my neighbors?
- Complexity of queries
 - Data is ingested from multiple sensors and manufacturers
 - Neighborhood data needs to be personalized for the notifications
- The data is updated as the real time data comes in. Eventual accuracy.

My challenges

- IoT data is largely real-time and needs to ingest as many requests as it can; Apache Kafka can handle >100K requests per sec vs RabbitMQ which can only handle >20K requests per sec
- Handle this data across 3,371,062 households just from NYC that's about 3000 Billion data points per day.
- Build infrastructure once and assure that it can handle the data blocks due to connection failures.
- Learn, research and implement the technology in two weeks effectively.

Other considerations

- A sample data of few megabytes is available and simulating few hundred gigabytes of real-time data based on the data formats from manufacturers like PTC.
- A cluster of four- server is required to deploy the stack that could process at the matching rate. Costs about \$68 per week on a m4.large.
- To validate the data engineering a known error cases are simulated to achieve the best results.

About me

Ajay Kavuri

[@pseudoaj](https://www.instagram.com/pseudoaj)

Programmer. Crawler. Happy soul.

