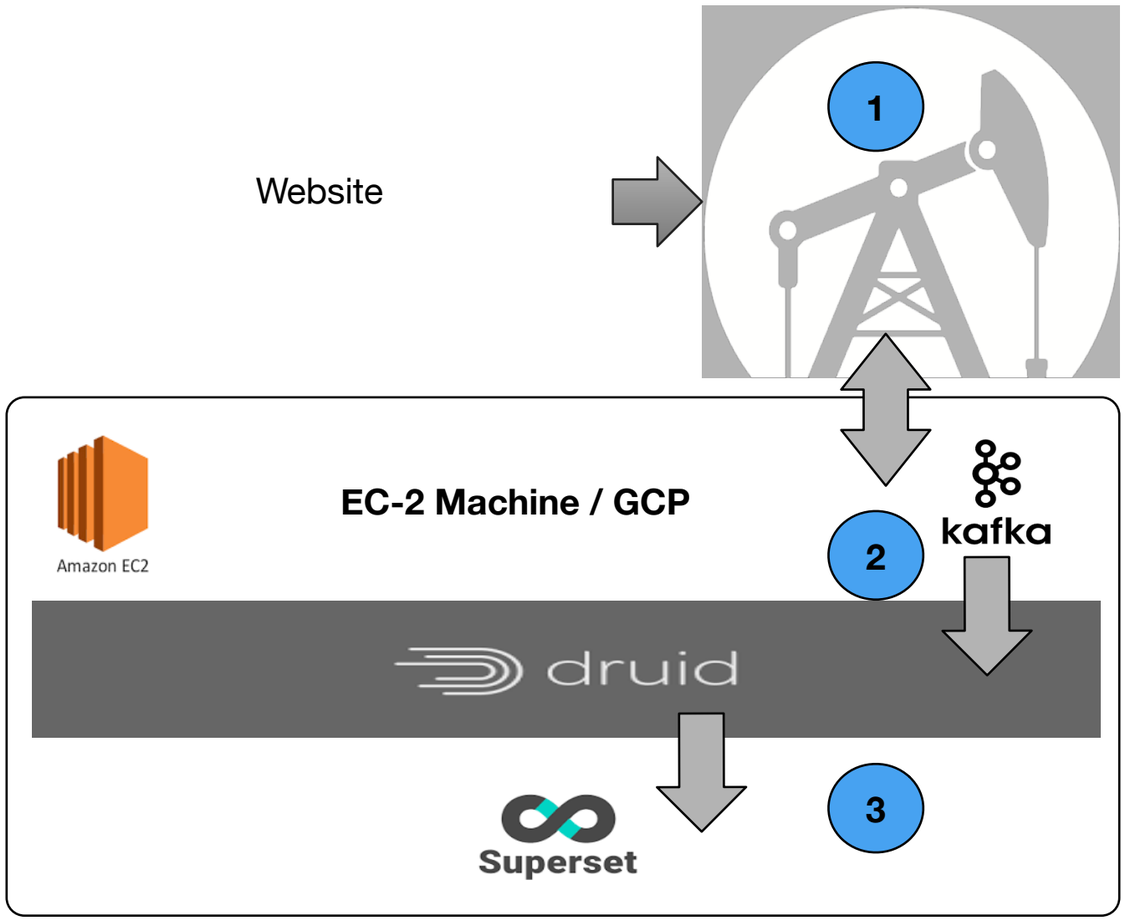
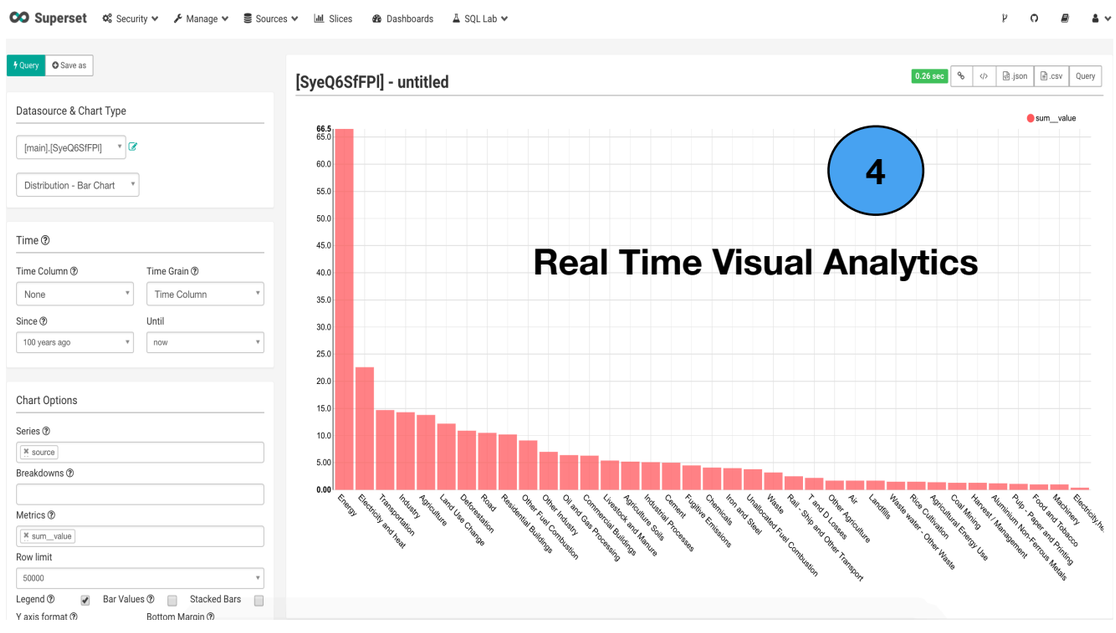
SCALABLE REAL-TIME DATA ANALYTICS ARCHITECTURE USING OPEN SOURCE TECHNOLOGIES





Learn what your customer wants as fast as possible , in near real time .

All the actions of users on your Website tells something about their intent. With the ability to process the data immediately, the business can tailor the content to each user.

1. Divolte (Data Collection)

Divolte Collector is a scalable and performant server for collecting clickstream data and publishing it to a sink, such as Kafka, HDFS or S3.
By using a small piece of JavaScript and a pixel in the browser of the customers, it gathers data about their behaviour on the website or application.

2. Kafka (Data Ingestion)

Apache™ Kafka is a fast, scalable, durable, and fault-tolerant publish-subscribe messaging system. Kafka is well known for its high throughput, reliability and replication. Kafka works well in combination with Apache Flink and Apache Spark for real-time analysis and rendering of streaming data.
In this setup Kafka is used to collect and buffer the events, that are then ingested by Druid. We need Kafka to persist the data and act as a buffer when there are bursts of events.

3. Druid (Data Storage in Batch and Realtime)

Druid is an open-source analytics data store designed for business intelligence (OLAP) queries on event data. Druid provides low latency real-time data ingestion from Kafka, flexible data exploration, and fast data aggregation.
Druid will do the processing of the data and shape it in the form that we request. Existing Druid deployments have scaled to trillions of events and petabytes of data, so we won't have to worry about scale.

4. Superset (Data Exploration)

Apache™ Superset is a data exploration and visualization web application and provides an intuitive interface to explore and visualize datasets, and create interactive dashboards. Initially developed by Airbnb, but now in the running to become an **Apache™ project**.