



Prof. Jeongkyu Lee

## Abstract

The diagram illustrates the Druid architecture components and their interactions:

- Streaming Data** and **Batch Data** are input sources.
- Real-time Nodes** (orange cylinder) receive streaming data and interact with **MySQL** (blue rectangle) and **Coordinator Nodes** (orange cylinder).
- MySQL** and **Coordinator Nodes** interact with **Zookeeper** (blue rectangle).
- Coordinator Nodes** and **Zookeeper** interact with **Broker Nodes** (orange cylinder).
- Batch Data** flows into **Deep Storage** (blue rectangle), which interacts with **Historical Nodes** (orange cylinder).
- Historical Nodes** interact with **Zookeeper** and **Broker Nodes**.
- Broker Nodes** serve **Client Queries**.
- Legend:**
  - Druid Nodes:** Orange cylinders (Real-time Nodes, Coordinator Nodes, Broker Nodes, Historical Nodes).
  - External Dependencies:** Blue rectangles (MySQL, Zookeeper, Deep Storage).
- Key:**
  - Queries:** Solid arrow.
  - Metadata:** Dotted arrow.
  - Data/Segments:** Dashed arrow.

Real-time nodes chunk data into segments, and are designed to frequently move these segments out to deep storage. To maintain cluster awareness of the location of data, these nodes must interact with MySQL to update metadata about the segments, and with Apache ZooKeeper to monitor their transfer.

## Results

The results for Wikipedia edits are as follows

The diagram illustrates a hybrid data architecture with the following components and data flow:

- Client:** Represented by a stack of documents on the left. It sends **DATA** (grey arrow) to the **BROKER** and receives **QUERIES** (red arrow) from the **BROKER**.
- Realtime:** Represented by green cubes at the top. It receives **DATA** (grey arrow) from the **BROKER** and sends **QUERIES** (red arrow) to the **BROKER**. It also receives **DATA** (grey arrow) from the **DATA STREAM** cloud and sends **QUERIES** (red arrow) to the **BROKER**.
- Broker:** Represented by orange cubes in the center. It acts as the central hub, receiving data from the Client, Realtime, and Data Stream, and sending queries to the Realtime and Historical components.
- Historical:** Represented by blue cubes on the right. It receives **DATA** (grey arrow) from the **BROKER** and sends **QUERIES** (red arrow) to the **BROKER**.
- Deep Storage:** Represented by a large blue cube at the bottom. It receives **DATA** (grey arrow) from the **BROKER** and sends **QUERIES** (red arrow) to the **BROKER**.
- Indexing:** Represented by horizontal bars. It is associated with the **DATA STREAM** and **BATCH DATA** clouds. The **INDEXING** process is shown as a grey arrow from the cloud to the **Realtime** component.

The diagram shows how queries and data flow through this architecture

