Druid: NoSQL based Analytics on Real-Time Data

Ujwal Sadhu
Prof. Jeongkyu Lee
Department of computer science
University of Bridgeport

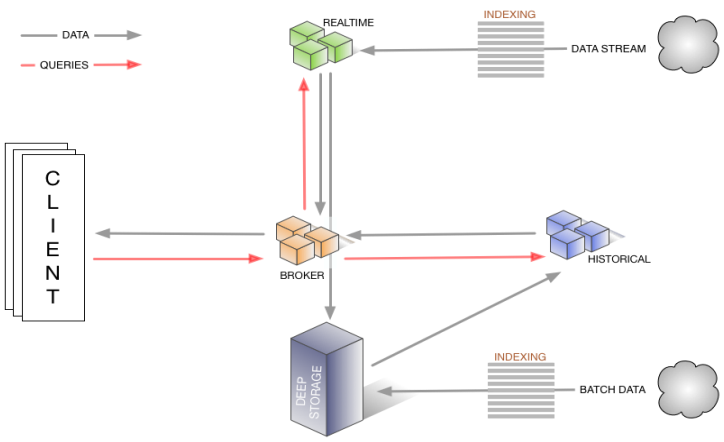
Abstract

Druid is an open source, distributed analytics data store that powers user-facing data applications, provides fast queries on data in Hadoop, and helps you glean insights from streaming data. Real-Time data is the data that is being generated in the current moment. The data generated by the Wikipedia edits or the Twitter tweets or the data from business can be used here.

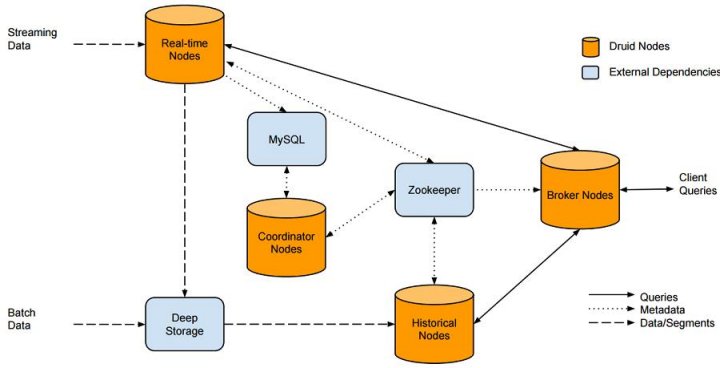
Introduction

Druid is an open source data store designed for real-time exploratory analytics on large data sets. The system combines a column-oriented storage layout, a distributed, shared-nothing architecture, and an advanced indexing structure to allow for the arbitrary exploration of billion-row tables with sub-second latencies. At a high-level, Druid collects event data into segments via real-time nodes. The real-time nodes push those segments into deep storage. Then a master node distributes those segments to compute nodes, which are capable of servicing queries. A broker node sits in front of everything and distributes queries to the right compute nodes.

Data Model



The diagram shows how queries and data flow through this architecture



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