

The Apache Hadoop Database for Fast Analytics on Fast Data

Random Reads/Writes

- Incorporate fast-changing, frequent updates into your data sets

Fast Analytic Scans

- Perform fast analysis on rapidly changing data

User-Friendly

- Simplify complex Lambda architectures into a single platform that can handle fast analytics on fast data

Integrated

- Take advantage of the broader Hadoop ecosystem—including Apache Spark, Apache Impala (incubating), and MapReduce

Open Source

- Fully ASF-governed open source project reduces user costs and increases portability

Designed for Next-Gen Hardware

- Kudu was developed in close collaboration with Intel, taking advantage of solid state drives and persistent memory

Apache Kudu™

Fast Analytics on Fast Data for Apache Hadoop™

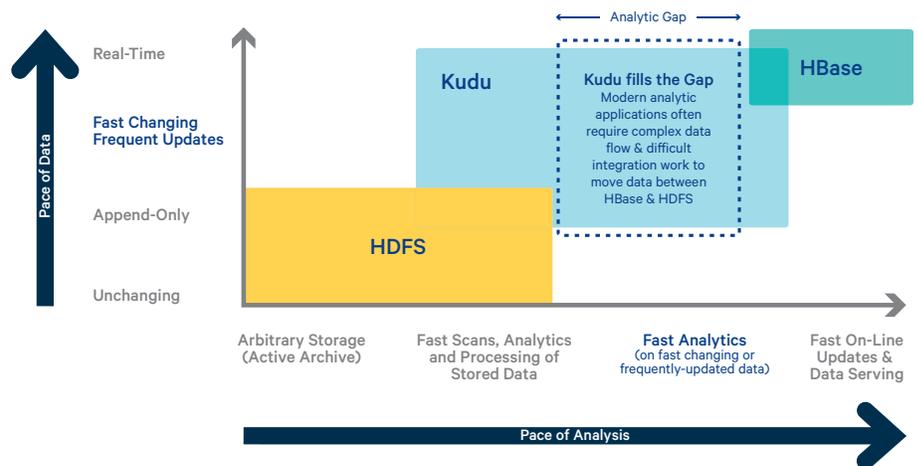
Apache Kudu is an open source storage layer that provides the Hadoop ecosystem with a database capable of fast analytics on fast data. Kudu eliminates the need for complex Lambda architectures that bifurcate data processing into speed and batch layers, instead combining them within one data layer. The simplification of the architecture opens up the Hadoop ecosystem as a low-cost alternative to a variety of use cases—including time series data, machine data analytics, online reporting, and predictive IoT.

Addressing the Gap in Hadoop Storage Layers

The original Hadoop storage layer is HDFS—Hadoop Distributed File System. HDFS enables the ingestion of large amounts of batch data, as well as the ability to analyze that data for trends and other insights. The capability to do fast analytic scans of this data is the signature of HDFS.

Apache HBase™ was introduced after HDFS to enable operational use cases that required the NoSQL-like capabilities of rapid reads and writes. HBase enables use cases where it's important to recall specific values and keep a real-time set of updated data available for specific queries.

Kudu fills the gap between these original Hadoop storage layers by combining the ability to rapidly ingest data with the ability to do broad analytic scans on this data. This native Hadoop storage engine provides a simplified architecture that leverages next-generation hardware. By removing the barrier of creating a complex Lambda architecture, Kudu makes it easier to launch new, real-time analytic applications, and enables the full update capabilities necessary for analytic workloads.



Kudu Use Cases

With Kudu, users can make analytic queries on real-time, updateable data. When combined with other components of the Apache Hadoop ecosystem, Kudu yields broad database solutions:

- **Analytic Database:** Together with Apache (incubating) Impala, Kudu creates a database that enables real-time analytics that allow users to take action on data while outcomes can still be swayed by informed decisions. Impala's leading analytic SQL performance, combined with Kudu's updateable columnar storage, provides a modern analytic database that brings high-performance BI and analytics to big data.
- **Operational Database:** Apache Spark™ integrates with Kudu to deliver real-time applications via Cloudera's operational database. Kudu together with Spark provide the ability for these applications to move beyond serving rule-based insights and data retrieval to higher-value analytical analysis that yields better answers.

In addition to enabling category database solutions, Kudu is particularly effective at addressing a specific set of use cases:

- **Time Series Data:** Time series data is most valuable if you can analyze it to change outcomes in real time. With Kudu, time series data can be inserted/updated as it arrives, analytic scans can find trends within fresh time series data, and random reads enable analysts to visit points in time where events occur.
- **Machine Data Analytics:** Kudu's ability to analyze trends through analytic scans, combined with real-time data inserts, means you can identify network problems before they happen. With Kudu, you can defend your network while ensuring its health and application performance.
- **Online Reporting:** Online reporting has typically been limited by data volume and analytic capacity, which meant that only recent data was available for granular query. Kudu changes this paradigm by enabling random reads/writes alongside analytics scans, transforming the capabilities of an operational data store.

Kudu in Cloudera's Enterprise Data Hub

Cloudera's development of Kudu and subsequent contribution to the Apache Software Foundation demonstrates a deep commitment to not only the project, but also to addressing new and changing customer use cases. Kudu's integration points with Impala, Spark, and MapReduce will continue to expand and include the full suite of components in the Cloudera Enterprise Data Hub, as well as the management, governance, and security features that users demand.

Download and try Kudu

<http://www.cloudera.com/downloads/beta/kudu.html>