HBase

HBase is a distributed column-oriented database built on top of HDFS. HBase is the Hadoop application to use when you require real-time read/write random access to very large datasets.

HBase comes at the scaling problem from the opposite direction. It is built from the ground up to scale linearly just by adding nodes. HBase is not relational and does not support SQL, but given the proper problem space, it is able to do what an RDBMS

cannot: host very large, sparsely populated tables on clusters made from commodity hardware.

The canonical HBase use case is the webtable, a table of crawled web pages and their attributes (such as language and MIME type) keyed by the web page URL. The webtable is large, with row counts that run into the billions. Batch analytic and parsing

MapReduce jobs are continuously run against the webtable, deriving statistics and adding new columns of verified MIME-type and parsed-text content for later indexing by a search engine. Concurrently, the table is randomly accessed by crawlers running

at various rates and updating random rows while random web pages are served in real time as users click on a website’s cached-page feature.

The HBase project was started toward the end of 2006 by Chad Walters and Jim

Kellerman at Powerset. It was modeled after Google’s “Bigtable:

**Types of NoSql Databases**

•Key Value Store

•A key and plain value

•Amazon Dynamo

•Document Store

•Stored as key and value, where the value can be a structure, like a JSON object. Each "document" can have all, some, or none of the same elements as another

•Mongo DB

•Column Families

•Each row (addressed by a key) contains one or more "columns". Columns are themselves key-value pairs. The column names need not be predefined, i.e. the structure isn't fixed.

•Columns are grouped as set of families and stored against a key

•Big Table, HBase