HBase Baseics for Baes

Jeremy Garcia, Sungwoo Park, Willem Thorbecke

What is HBase?

- NoSQL key value data store
- Built on top of HDFS
  - Hadoop distributed file system
- Based on Google’s BigTable: A Distributed Storage System for Structured Data

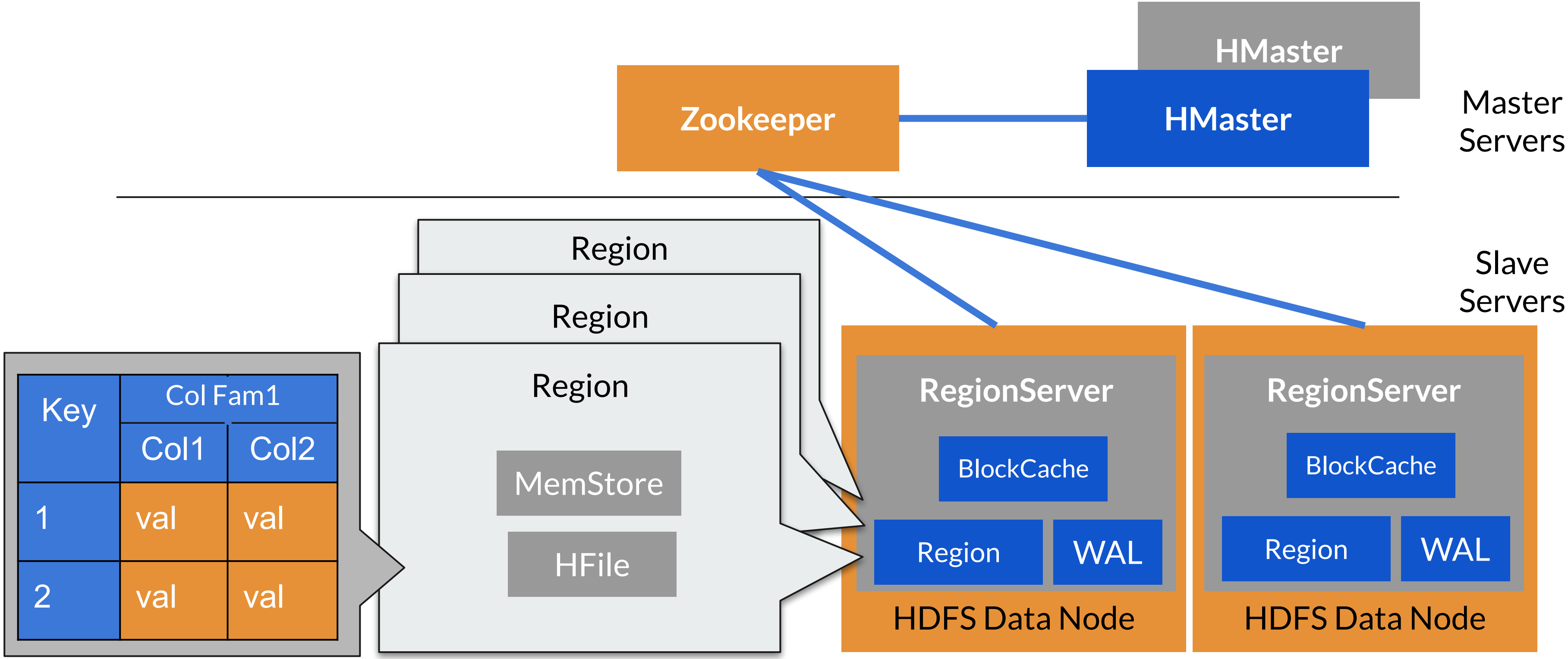
When to use HBase?

- Processing large data (PB/TB range)
- Operations such as data reading and processing will take small amount of time compared to traditional relational models
- Random read/write access is needed for Big Data

HBase versus RDBMS

(Sort of) schema-less in database	Governed by a schema
Column-oriented	Row-oriented
Wide and sparsely populated tables. Horizontally scalable	Thin and built for small tables. Hard to scale
Designed to store de-normalized data	Has normalized data
Supports automatic partitioning	No built in support for partitioning
Enables aggregation over many rows and columns	Aggregation is an expensive operation

HBase Architecture



Key	Col Fam1	
	Col1	Col2
1	val	val
2	val	val

Applying HBase to Basketball Analytics

- We decided to use HBase to make a real-time winning probability engine for basketball games
- We use HBase to store play-by-play data for each game played in the past 10 years (12,300 games)
    - Scale does not quite utilizes HBases potential (but close)
- Redundancy and reliably of Hadoop, not currently utilized (all data stored locally)
- MVP
    - Throughout the course of a live game, predict outcome by referencing all other games with the same score differential at the same point in time in the game
- Stretch
    - Use more variables to narrow the definition of a similar game on top of a score and time in game
        - E.g., make sure the team's have similar records, who has possession at the end of the game
        - Use linear regression to figure out what stats are correlated the most with the outcome of a game

Example

(Away Team) 76 points: 68 points (Home Team)

38 seconds left on 4th quarter

We look at games in the past where home team is down by 8 points with 38 +/- 3 seconds left.

We then calculate in how many of those games a home team won.

Implementation

- We spun up a local instance of HBase on one of our laptops
- Using the Python library HappyBase we populated HBase with minimal version of our NBA data set

Time	Quarter	Score	Score Margin	Result
0:38	4	76-68	-8	Home

- Access all other games that have this same score margin (-8), ignore all games that are at a different time, use Result column to calculate odds of winning

Thanks for listening!

Any questions?