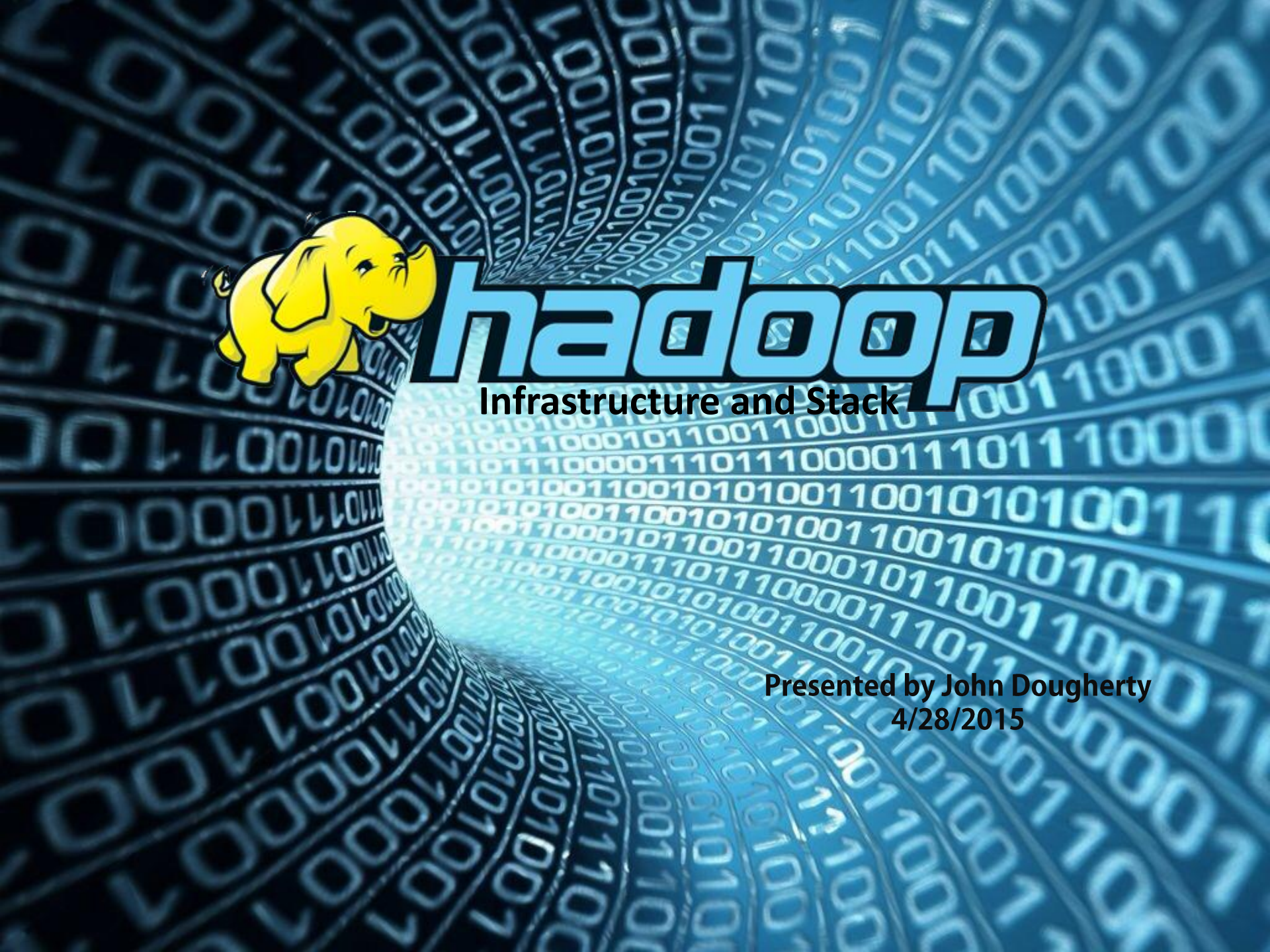


Infrastructure and Stack

Presented by John Dougherty

4/28/2015

What is Hadoop?

- Apache's implementation of Google's BigTable
- Uses a Java VM in order to parse instructions
- Uses sequential writes & column based file structures with HDFS
- Grants the ability to read/write/manipulate very large data sets/structures.

What is Hadoop? (cont.)

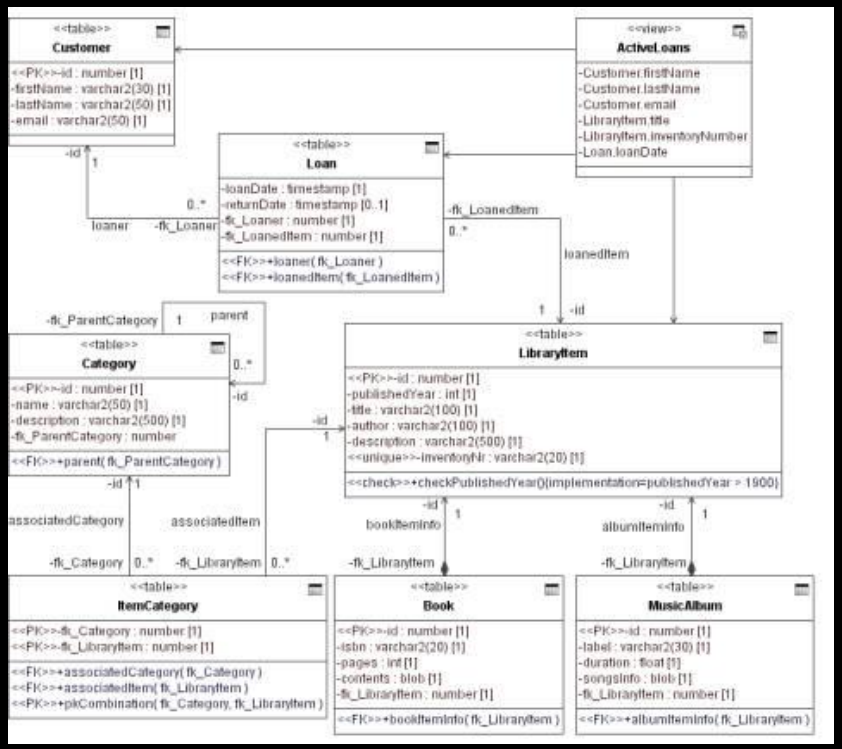
C1 :C1

store
recall
A B
C D
(*
0 -
ce +
enter

Resize
Fill
Generate
Iterate
Transfo...
SortAsc...
SortDes...
Maximum
Minimum
Average
SD : N-1
SD : N
Correlat...
Σ?
Σ?^2
Σxy
N

	x	y
R1	9.424777961	0.048000000
R2	10.21017612	0.065115282
R3	10.99557429	0.072000000
R4	11.78097245	0.096166522
R5	12.56637061	0.112000000
R6	13.35176878	0.132966161
R7	14.13716694	0.140000000
R8	14.92256510	0.158391919
R9	15.70796327	0.180000000
R10	16.49336143	0.200838243
R11	17.27875959	0.228000000
R12	18.06415776	0.265872150
R13	18.84955592	0.324000000
R14	19.63495408	0.367695526
R15	20.42035225	0.404000000
R16	21.20575041	0.463862048
R17	21.99114858	0.540000000
R18	22.77654674	0.622253967
R19	23.56194490	0.692000000
R20	24.34734307	0.814587012
R21	25.13274123	0.976000000

VS.



BIG DATA

What is BigTable

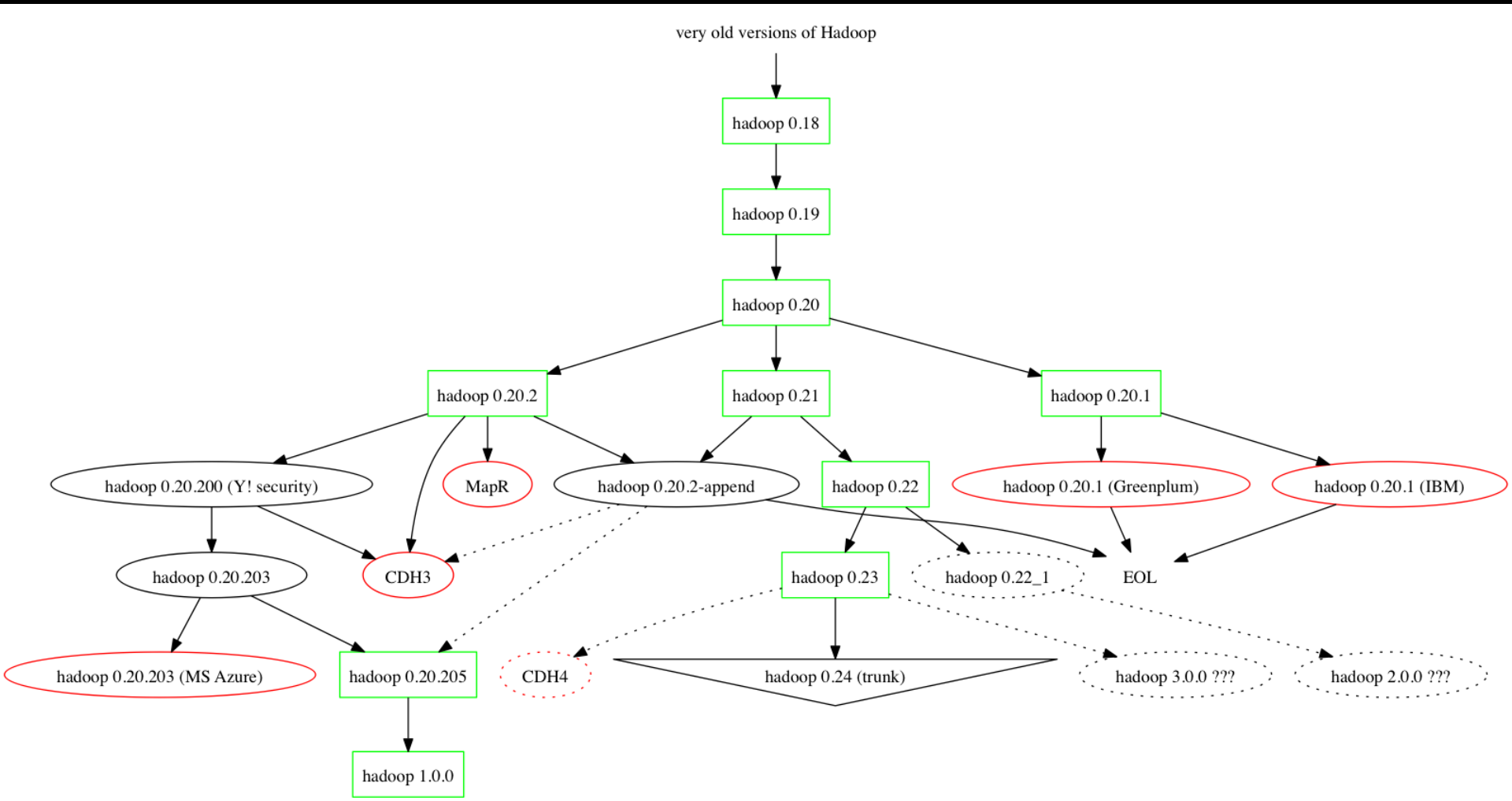
- Contains the framework that was based on, and is used in, hadoop
- Uses a commodity approach to hardware
- Extreme scalability and redundancy
- Is a compressed, high performance data storage system built on Google's File System

Commodity Perspective

- Commercial Hardware cost vs. failure rate
  - Roughly double the cost of commodity
  - Roughly 5% failure rate

- Commodity Hardware cost vs. failure rate
  - Roughly half the cost of commercial
  - Rougly 10-15% failure rate

Breaking Down the Complexity



What is HDFS

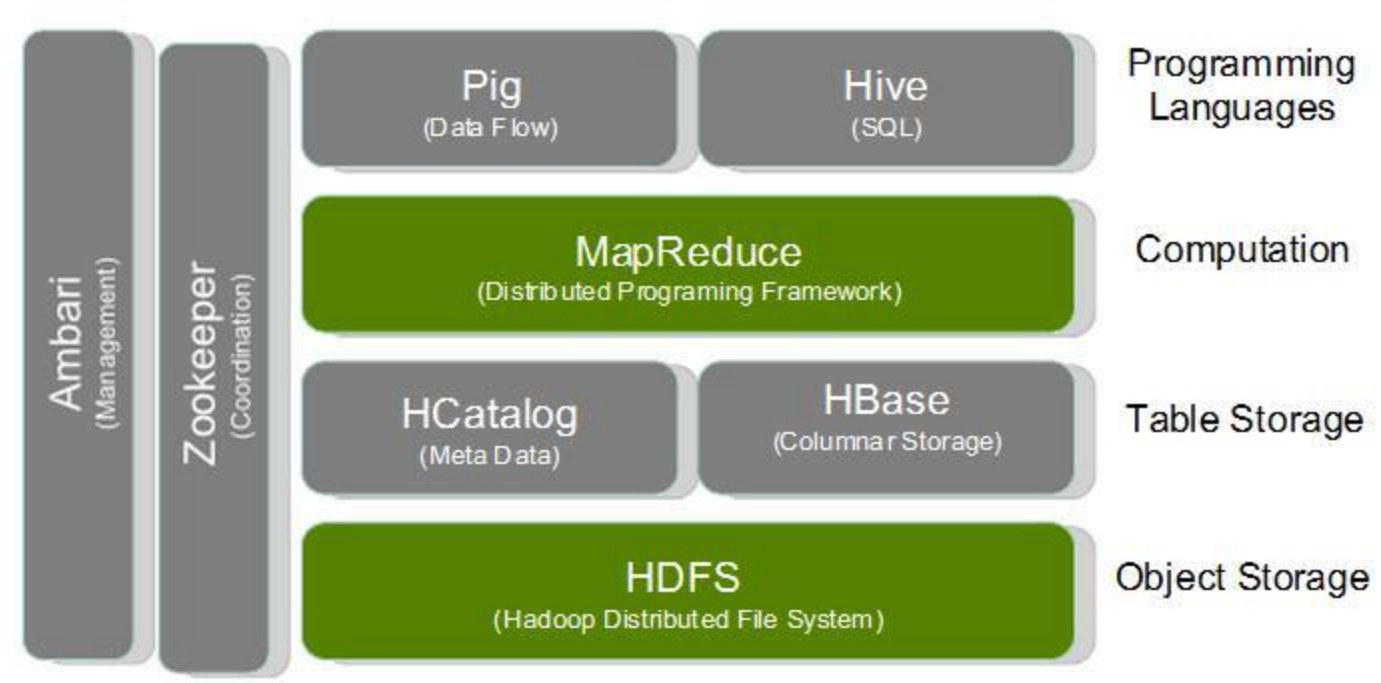
- Backend file system for the Hadoop platform
- Allows for easy operability/node management
- Certain technologies can replace or augment
  - Hbase (Augments HDFS)
  - Cassandra (Replaces HDFS)

What works with Hadoop?

- Middleware and connectivity tools improve functionality
- Hive, Pig, Cassandra (all sub-projects of Apache's Hadoop) help to connect and utilize
- Each application set has different uses

Pig

Layout of Middleware



Schedulers/Configurators

- Zookeeper
  - Helps you in configuring many nodes
  - Can be integrated easily
- Oozie
  - A job resource/scheduler for hadoop
  - Open source
- Flume
  - Concatenator/Aggregator (Dist. log collection)

Middleware

- Hive
  - Data warehouse, connects natively to hadoop's internals
  - Uses HiveQL to create queries
  - Easily extendable with plugins/macros
- Pig
  - Hive-like in that it uses its own query language (pig latin)
  - Easily extendable, more like SQL than Hive
- Sqoop
  - Connects databases and datasets
  - Limited, but powerful

How can Hadoop/Hbase/MapReduce help?

- You have a very large data set(s)
- You require results on your data in a timely manner
- You don’t enjoy spending millions on infrastructure
- Your data is large enough to cause a classic RDBMS headaches

Column Based Data

Developer woes

- Extract/Transfer/Load is still a concern for complicated schemas

- Egress/Ingress between existing queries/results becomes complicated

- Solutions are deployed with walls of functionality

- Hard questions turn into hard queries

Column Based Data (cont.)

Developer joys

- You can now process PB, into EB, and beyond
- Your extended datasets can be aggregated, not easily; but also unlike ever before
- You can extend your daily queries to include historical data, even incorporating into existing real-time data usage

Future Projects/Approaches

- Cross discipline data sharing/comparisons
- Complex statistical models re-constructed
- Massive data set conglomeration and standardization (Public sector data, etc.)

How some software makes it easier

- Alteryx
  - Very similar to Talend for interface, visual
  - Allows easy integration into reporting (Crystal Reports)
- Qubole
  - This will be expanded on shortly
  - Easy to use interface and management of data
- Hortonworks (Open Source)
  - Management utility for internal cluster deployments
- Cloudera (Open, to an extent)
  - Management utility from Cloudera, also for internal deployments