Hadoop: Scalable Infrastructure for Big Data

QCon London 2012

Parand Tony Darugar
Founder and CEO, Xpenser
parand@xpenser.com

What is Hadoop?

Hadoop is the Linux of Big Data Processing

Infrastructure
for
Large Scale Computation
& Data Processing
on a network of
Commodity Hardware.

Why Hadoop?

Scale

Cost

Freedom

Does Anyone Use Hadoop?

IBM	eHarmony
VISA	Zion's bank
Microsoft	NY Times
Facebook	Twitter
Yahoo	eBay
AOL	LinkedIn
…	…

Alternatives

Build your own

Get creative with RDBMS architecture

What's the idea?

Commodity Hardware

Distributed Operation

Wisdom:

Embrace Failure
(hardware)

Be Resilient
(software)

What's in the box?

Hadoop Distributed File System

Distributed Computation Framework

Map-Reduce Programming Model

HDFS

- Your data in triplicate
- Built-in resiliency to large scale failures
- Intelligent Data Distribution
- Very large data sizes

Distributed Computation

- Built-in resiliency to large scale failures
- Distribute work to workers, collect results from fastest
- Move computation to data (not data to computation)

Map Reduce

Very simple programming model:

Map(anything)->key, value

Sort, partition on key

Reduce(key,value)->key, value

No parallel processing or message passing semantics

Programmable in Java or any other language (streaming)

Ecosystem

HBase: NoSQL BigTable clone

Hive: Somewhat-SQL data store

Pig: SQL-like programming model

Chukwa, Scribe, Mahoot, Cassandra, Oozie, Sqoop, ...

Commercial Support

Cloudera

HortonWorks

IBM

...

How?

Try it in non-distributed mode

Try it on a few spare machines

Try it on EC2

Try it! http://hadoop.apache.org/

Case Studies

eHarmony

Biz360 (Attensity)

Yahoo!

You!

Start with ETL

Start with batch, non time-critical tasks

Start with storing your large data on HDFS

Move batch processing to Hadoop

Serve from RDBMS

Embrace. Be One With The Hadoop.

Questions?

Parand Tony Darugar
parand@xpenser.com