MAPR TECHNOLOGIES

EASY. DEPENDABLE. FAST.

Hadoop

The present to the Future

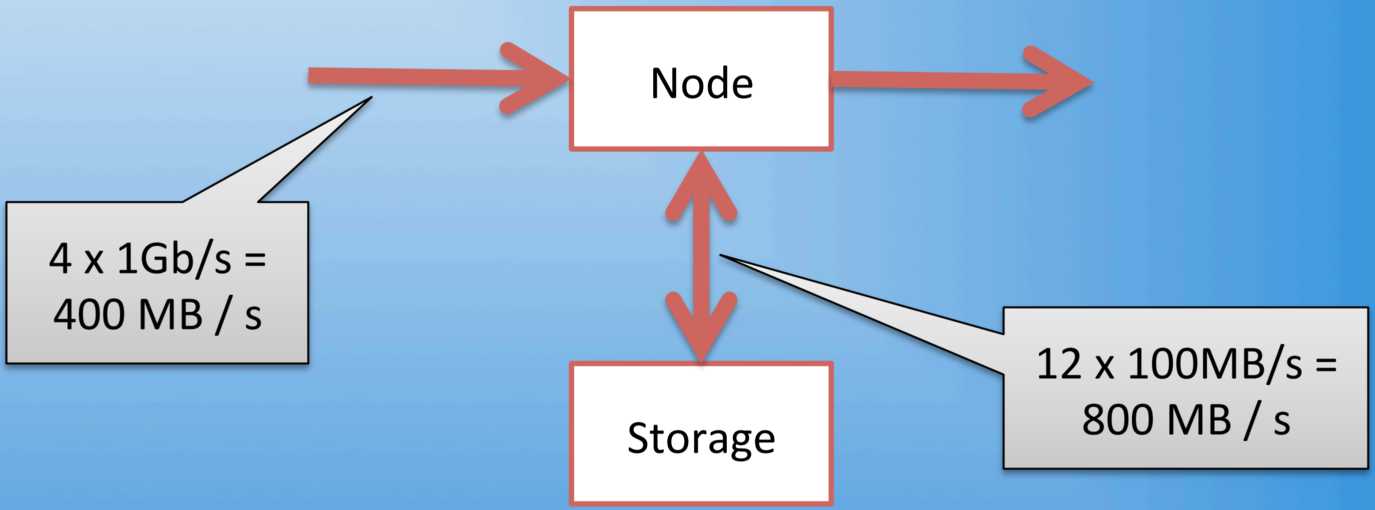
Some Notes

- Tweet questions and comments with #bbuzzted
- Take and share notes in real-time at http://tinyurl.com/buzzwords-ted-dunning

Where are we now?

- Technically, well below potential
  - Framework moves data at 1/3 … 1/10 potential
    - -0.5 to -1 on S scale
  - Framework small programs at 1/10 … 1/00
    - -1 to -2 on S scale
  - Flow based programs not supported
  - Integration into larger data flows severely hampered

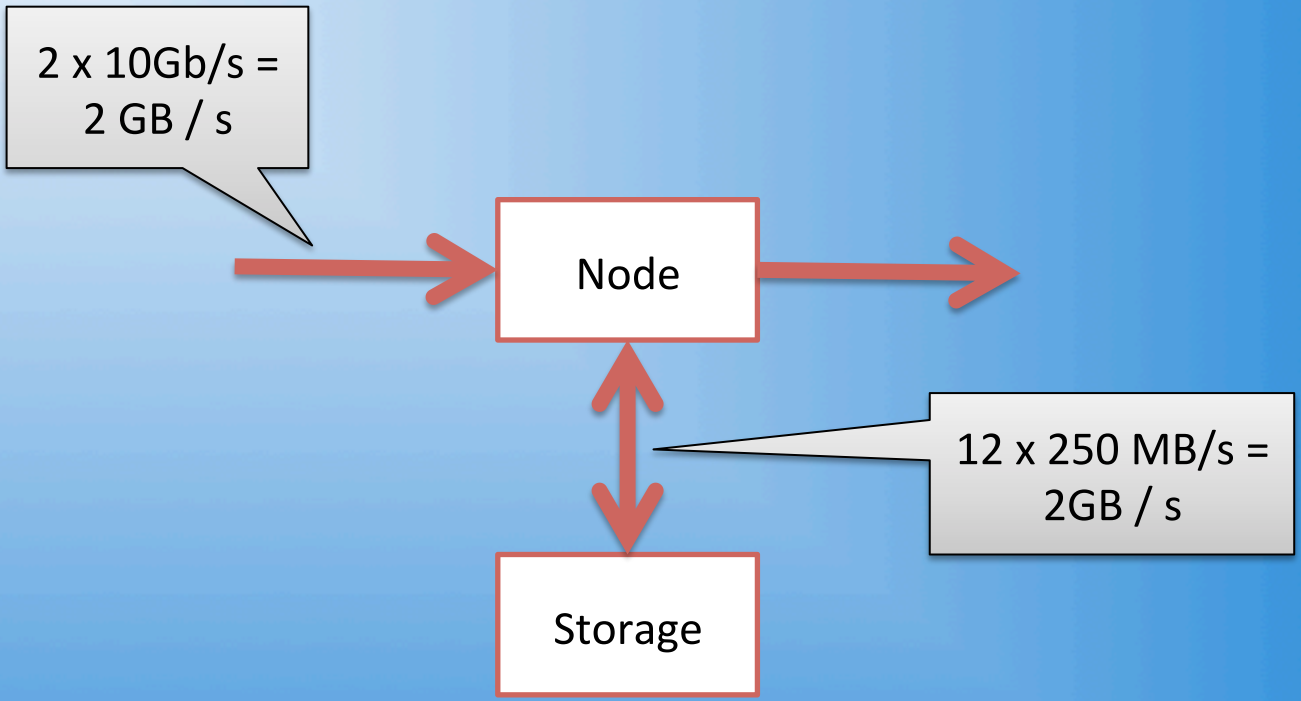
Data Flow Expected Volumes



4 x 1Gb/s =
400 MB / s

12 x 100MB/s =
800 MB / s

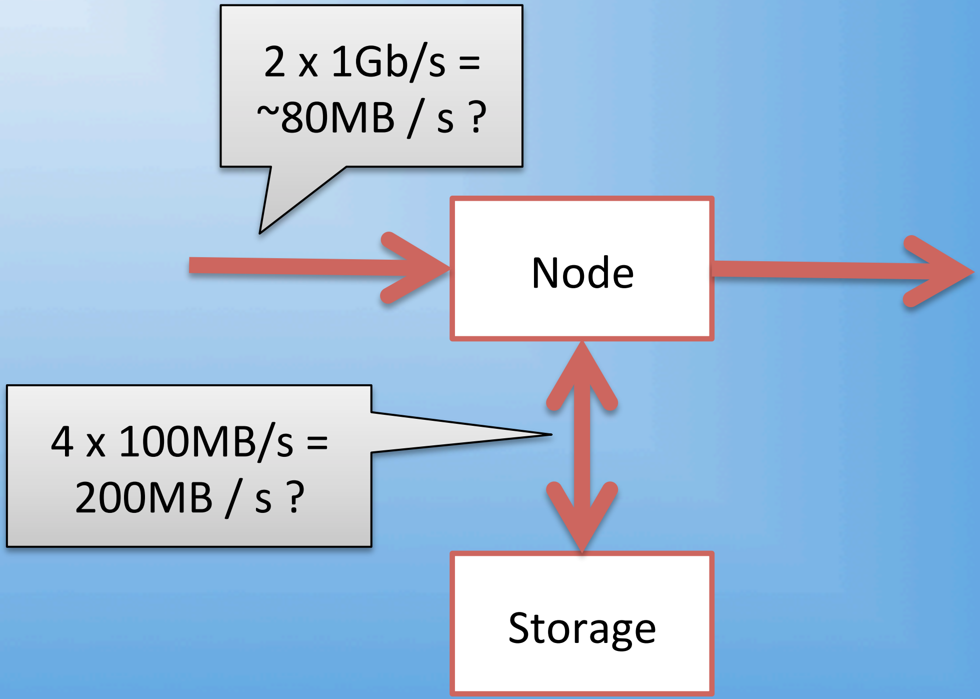
Data Flow Potential Volumes



2 x 10Gb/s =
2 GB / s

12 x 250 MB/s =
2GB / s

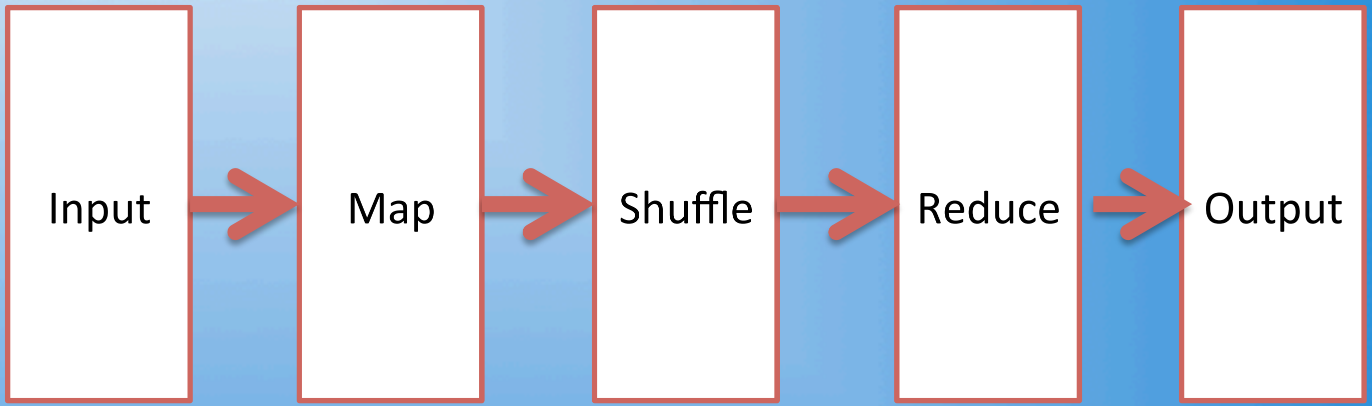
Data Flow Current Hadoop Volumes



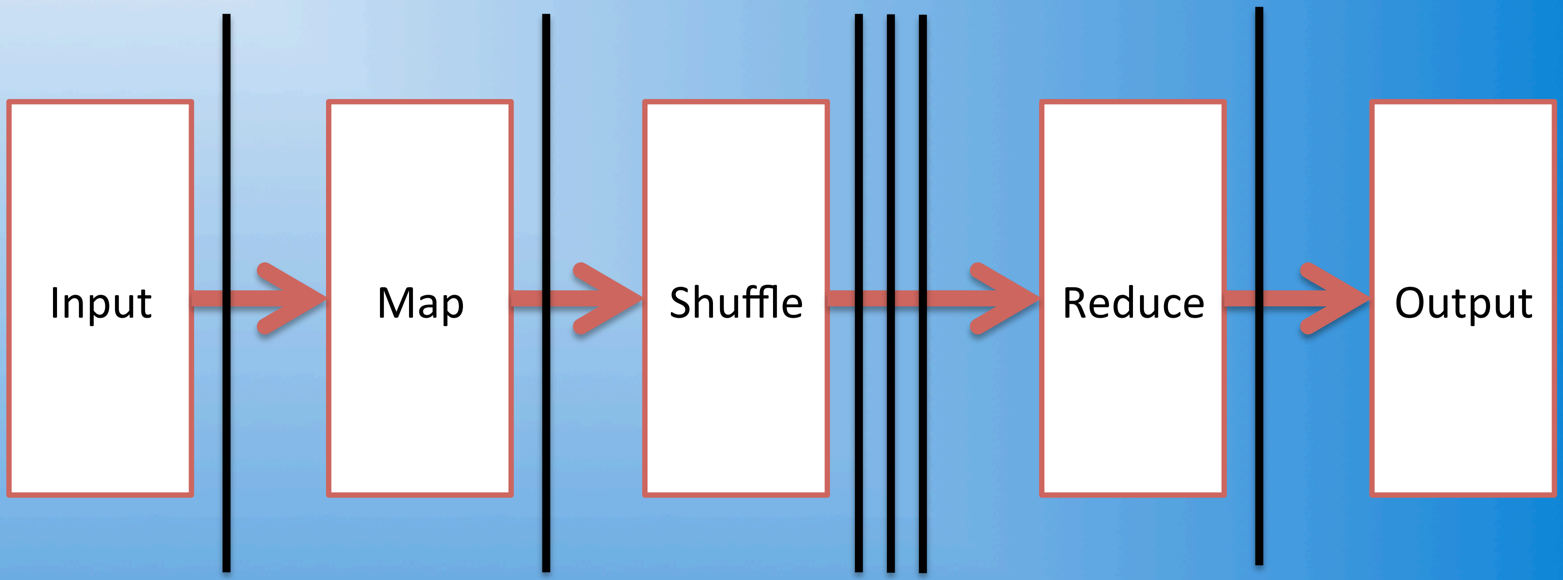
2 x 1Gb/s = ~80MB / s ?

4 x 100MB/s = 200MB / s ?

What Happens?



What Actually Happens?



But Technical Problems are “easy”

- MapR solution cost us ~ 20 person-years
  - (not just any person)
- MR 2.0 is coming (really)

A Provocateur's History

- Nutch had a problem
- A small team produced a quick solution
- Lots of others piled on
- Things got out of hand
- Investors got involved
- And then … beer on the train

Hadoop's (social) Problems

- Community as a virtue
  - User satisfaction is self-fulfilling
- Community as a bug
  - Assumes a single agenda
  - Assumes self-scratching
- Serious money now involved
- No single agenda any more
- Lots of players building products

What is Community?

- Shared values (merit, transparency)
- Shared goals (openness, quality, self-service)
- Focused point of contact
- Long-term personal contact
- Members are individuals
- Sense of trust

What is an Eco-system?

- Multiple agendas
- Many corporate and individual entities
- Explicitly different sets of values
- Explicitly conflicting goals
- Direct competition
- Apache community is one piece of the eco-system

The Darker Road

- If we pretend an eco-system is a community, then:
  - Commit wars continue
  - Factionalization dominates
  - 101 forks
  - No compatibility, no way to test compatibility
  - Azure takes over

The Brighter Path

- A thousand flowers bloom
- All rooted in a common core
- Apache becomes one player among many
  - Apache Hadoop becomes the reference
  - Community => eco-system
- Users get a more predictable world
- Innovation welcomed
- Multiple development models flourish
- Many (very different) winners

But what kind of eco-system?

- Choose one:
- A totally mercenary, totally lethal, totally private eco-system?
  - What company does this best? (monkey boy!)
  - How many winners will there be?
- Or a very human, collaborative eco-system?
  - Do companies dominate this?

The key task

- The difference is whether the humans involved find common ground

Alternative 1: The Big Crash

- Inverse to the big bang
- Positive curvature universe, time-like paths are closed
- The universe expands and then collapses back into a Massive Singularity

Alternative 2: Inflationary Multiverse

- The universe (market) expands in a hyperbolic way
- Points separated by more than trivial distances are in separate light-cones
- The universe becomes a cold lonely place
- Critical mass is lost
- Can you say CPM?

Alternative 3: Green Software

- Zones of common interest (where possible)
- We build a (little bit of) formal structure
  - Call it a Consortium
  - Apache should participate, but is not the right structure
- All parties need to recognize and value alternative view points
  - Companies need to remember where the golden egg came from and why and how
  - Individuals need to see the power of concerted action
- Universal domination (ish)

What do we have now?

- Open source
- Commercial tools
- Commercial support
- Radical Surgery
- Major corporate usage

Open source

- From Apache:
  - Hadoop, Zookeeper, Lucene, Hive, Mahout, Pig, Hbase, Flume, Avro
- From others:
  - Plume, Mesos, Twister, Spark

Commercial tools

- DataMeer
- Karmasphere
- Kitenga
- Cascading

Support and Professional Services

- Cloudera
- Impetus
- ThinkBig Analytics
- Digital Reasoning
- Booz Allen
- IBM

Radical Surgery

- MapR
- DataStacks
- Spark

Major Users

- Major corporate usage
  - Twitter, Facebook, Yahoo

- Others whose names cannot be said

- And a horde of leprechauns

Conflicts, Real and Imagined

- Agility versus Stability
- Append versus No-Way
- Local Write Path versus No-Way
- Major Forklift versus Only-Apache
- Proprietary versus Open Source
- What is Apache

Common Ground

- An expanding universe is good
- Allowing innovation is good
- Supporting innovation is good
- Inspiring innovation is good
- Supporting different uses is good

Which Future Do We Want?

- Can't we all just get along?

Which Future Do We Want?

- Can’t we all just get along?
  - No. Probably not
- Well, can we get along when we agree?
- … Or when it doesn’t really matter?
- … Or where we could both move to an alternative acceptable to both?

Inevitabilities

- We have an eco-system
- It can be functional
- Or disfunctional

Options

- We can make our eco-system work
- It isn't what it was
- And it never will be
- But it can be astonishing

Thank You!

Ted Dunning

tdunning@maprtech.com

tdunning@apache.com

http://tdunning.blogspot.com/

@ted_dunning

Comments!

- Tweet questions and comments with

  #bbuzzted

- Take and share notes in real-time at

  http://tinyurl.com/buzzwords-ted-dunning