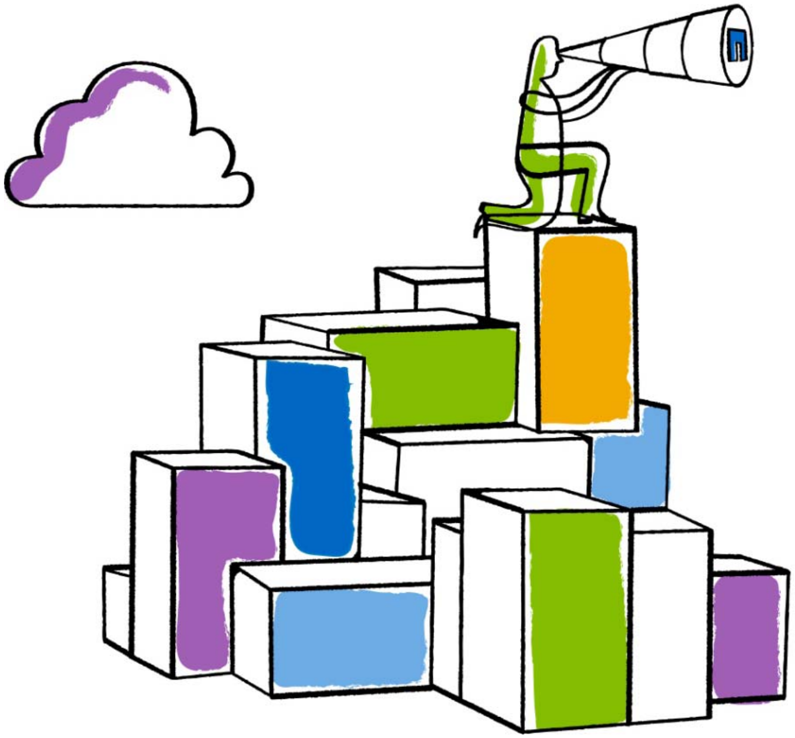
Go further, faster®



# Big Data – Trends to Watch

Bill Peterson

NetApp

September, 2012

© 2012 NetApp, Inc. All rights reserved.

**Bill Peterson**

@thebillp

# What I hope to accomplish today...

Think beyond structured and unstructured data

&

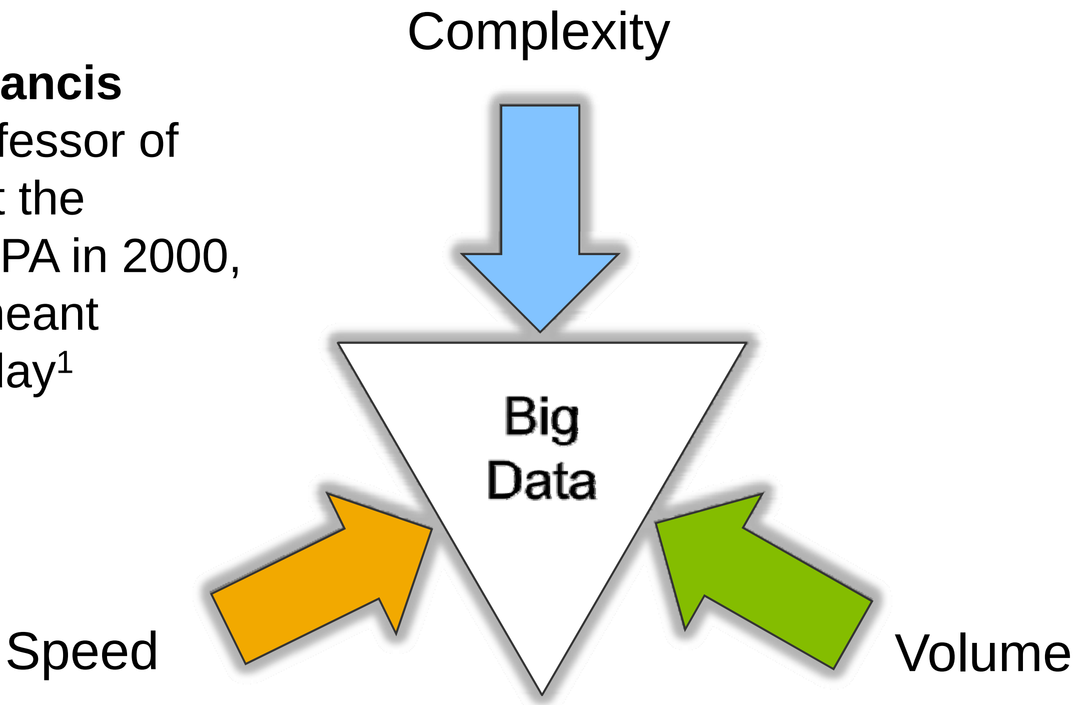
Think beyond big data hype

...and avoid this.

# What is “Big Data”?

“Big Data” refers to datasets whose volume, speed and complexity is beyond the ability of typical tools to capture, store, manage and analyze.

Coined by **Francis Diebold**, professor of economics at the University of PA in 2000, when “Big” meant Gigabytes / day[1]



© 2012 NetApp, Inc. All rights reserved.

# Quantifying The Big Data Challenge

## 5 Billion

smart
phones

## 60 Zettabytes

Estimated size of the digital universe in 2020

## Growth Over the Next Decade:

| | |
|---|---|
| Servers (Phys/VM): | 10x |
| Data/Information: | 50x |
| #Files: | 75x |
| IT Professionals: | <1.5x |

**Source:** *Gantz, John and Reinsel, David,* "Extracting Value from Chaos", IDC IVIEW, *June 2011, page 4.*

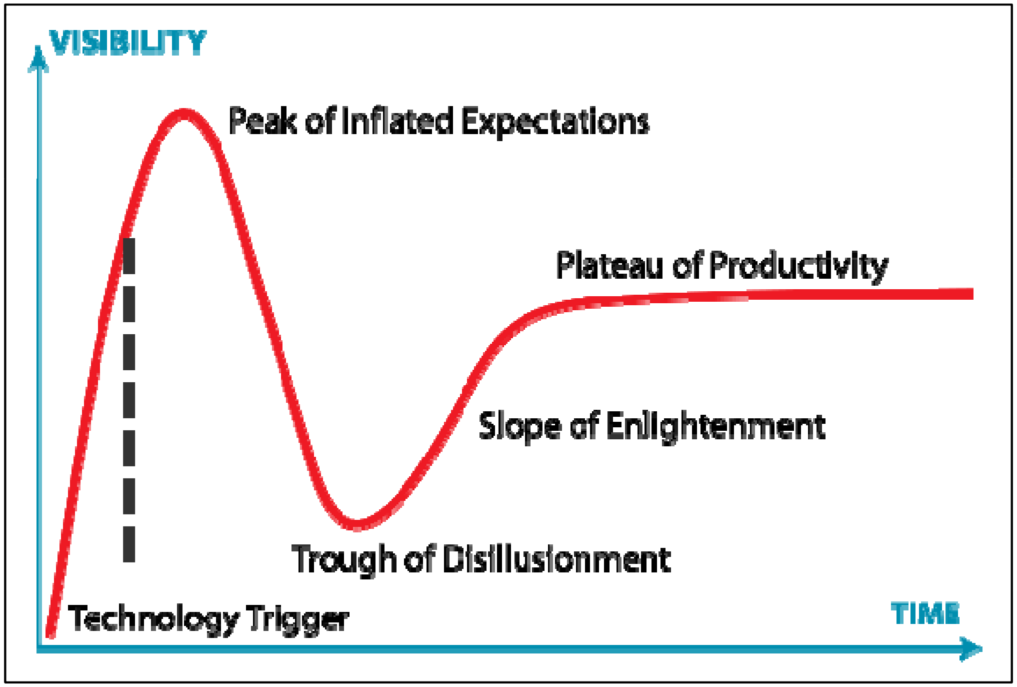
## 30 Billion

pieces of new content to Facebook per month

Sensors
Video
Music
Location
Weblogs

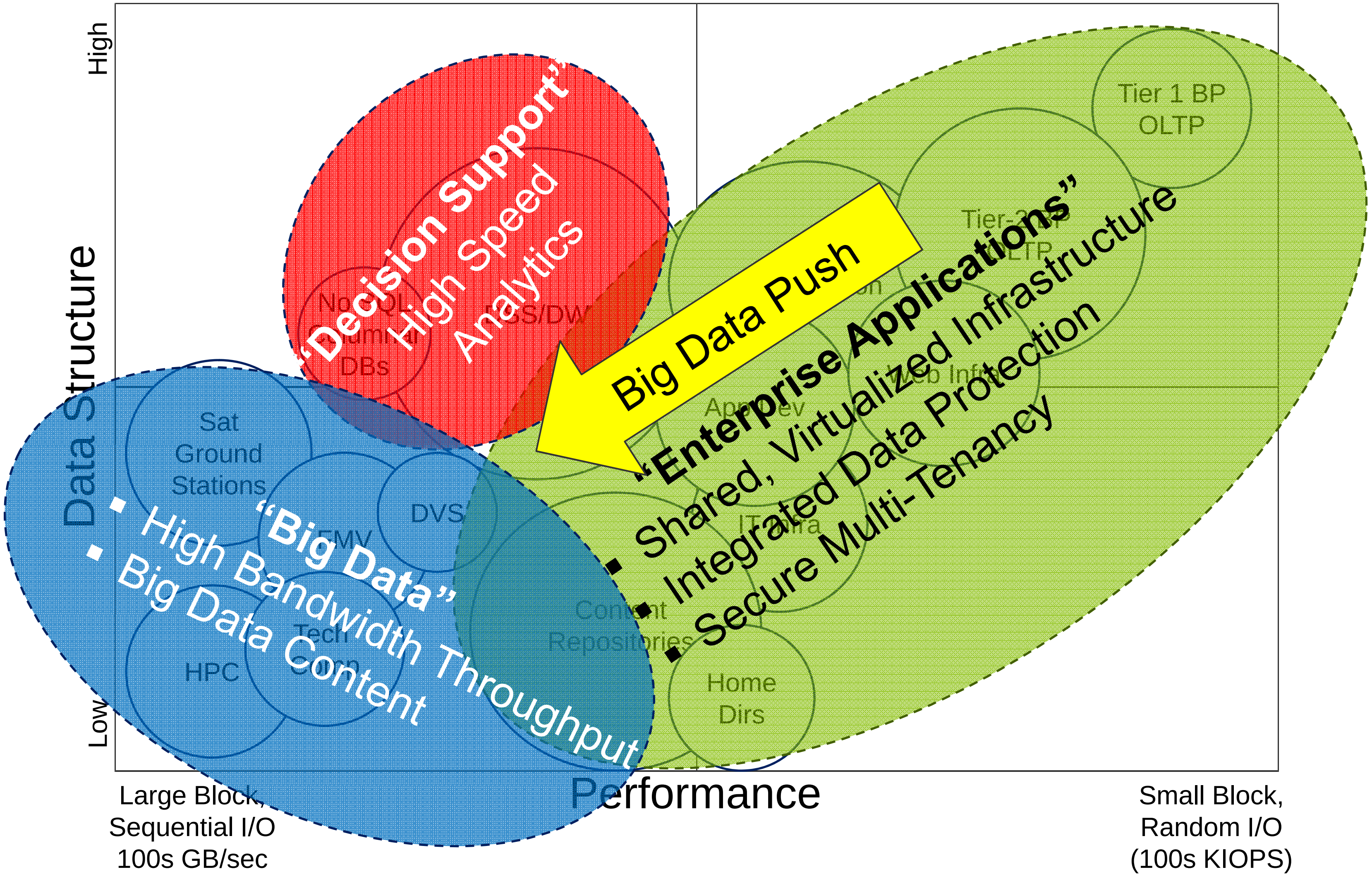
**80%**

of data is unstructured



© 2012 NetApp, Inc. All rights reserved.

# The Big Data Push



“Decision Support”
High Speed Analytics

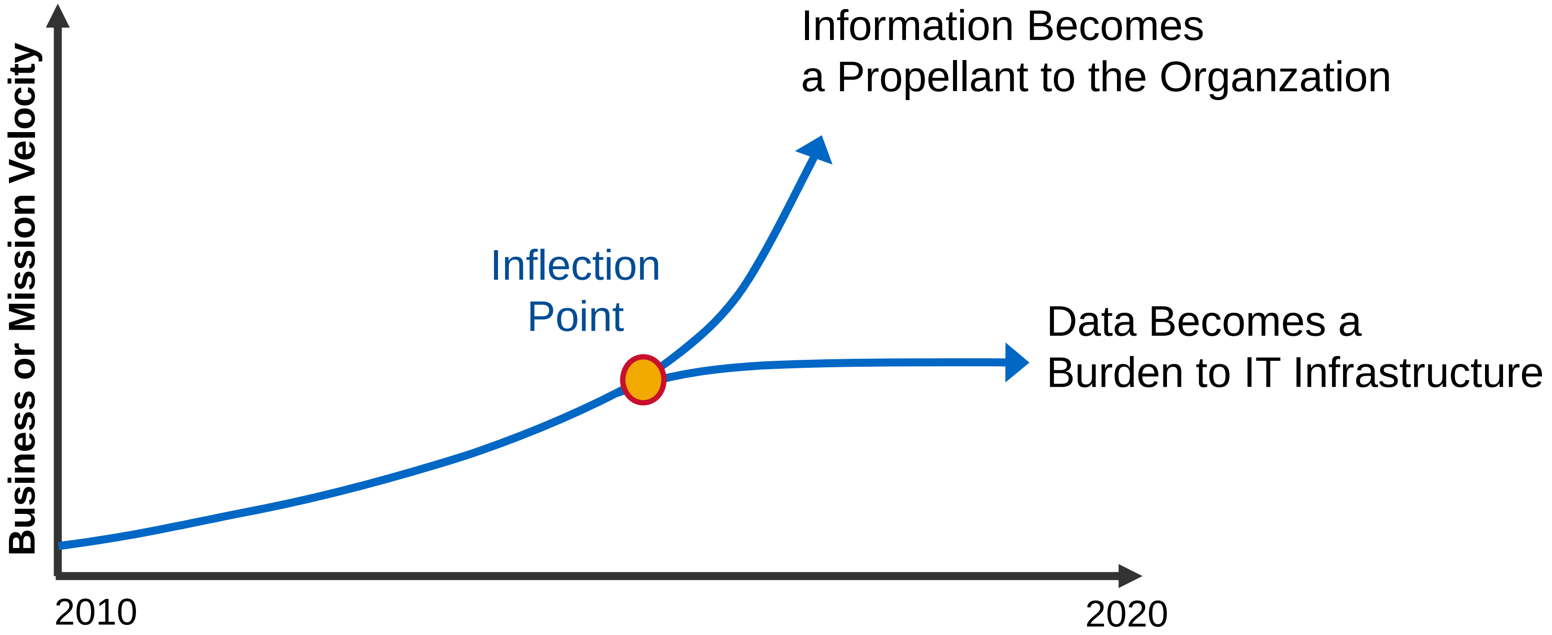
“Big Data”
- High Bandwidth Throughput
- Big Data Content

Big Data Push

“Enterprise Applications”
- Shared, Virtualized Infrastructure
- Integrated Data Protection
- Secure Multi-Tenancy

© 2012 NetApp, Inc. All rights reserved.

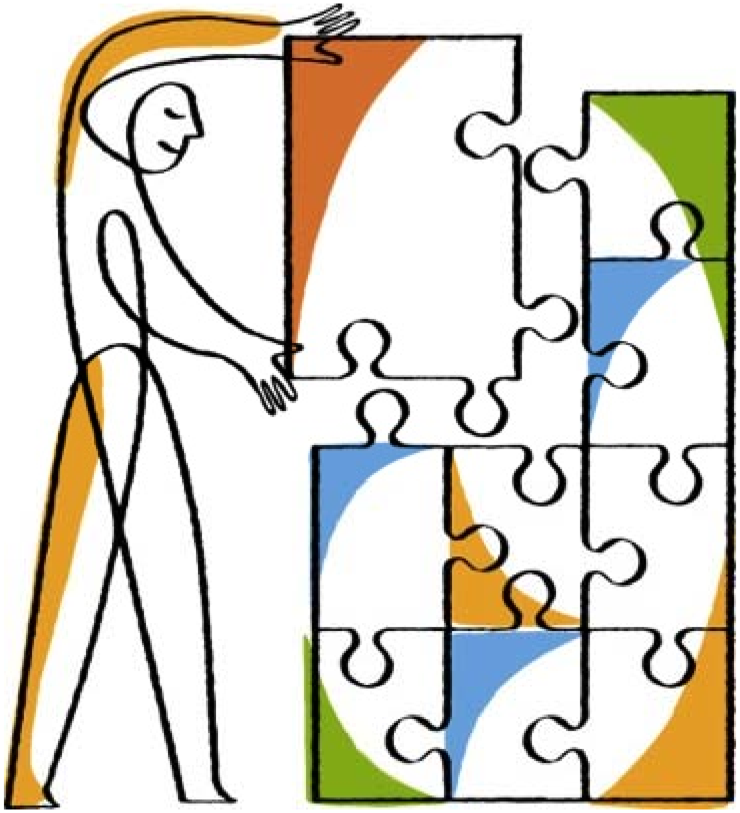
# What Does This Mean to You?



You are also at an Inflection Point:  You also have a decision to make, as “business as usual” may not cut it!

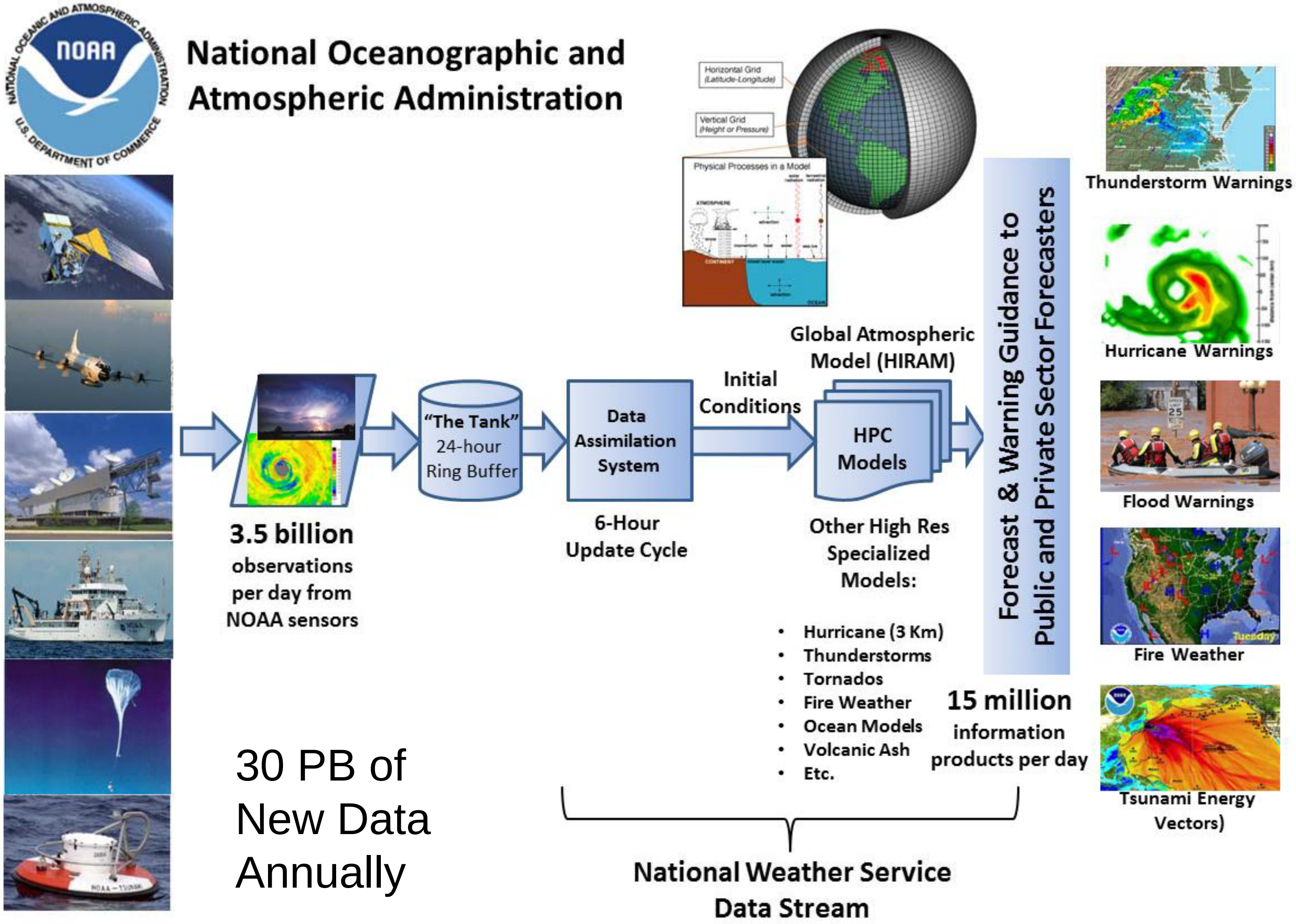
© 2012 NetApp, Inc. All rights reserved.

# Dispelling the Misconceptions About Big Data



© 2012 NetApp, Inc. All rights reserved.

# Big Data Is NOT New



National Oceanographic and Atmospheric Administration

Horizontal Grid (Latitude-Longitude)
Vertical Grid (Height or Pressure)
Physical Processes in a Model

3.5 billion observations per day from NOAA sensors

"The Tank" 24-hour Ring Buffer

Data Assimilation System

6-Hour Update Cycle

Initial Conditions

Global Atmospheric Model (HIRAM)

HPC Models

Other High Res Specialized Models:

- Hurricane (3 Km)
- Thunderstorms
- Tornados
- Fire Weather
- Ocean Models
- Volcanic Ash
- Etc.

Forecast & Warning Guidance to Public and Private Sector Forecasters

15 million information products per day

Thunderstorm Warnings

Hurricane Warnings

Flood Warnings

Fire Weather

Tsunami Energy Vectors)

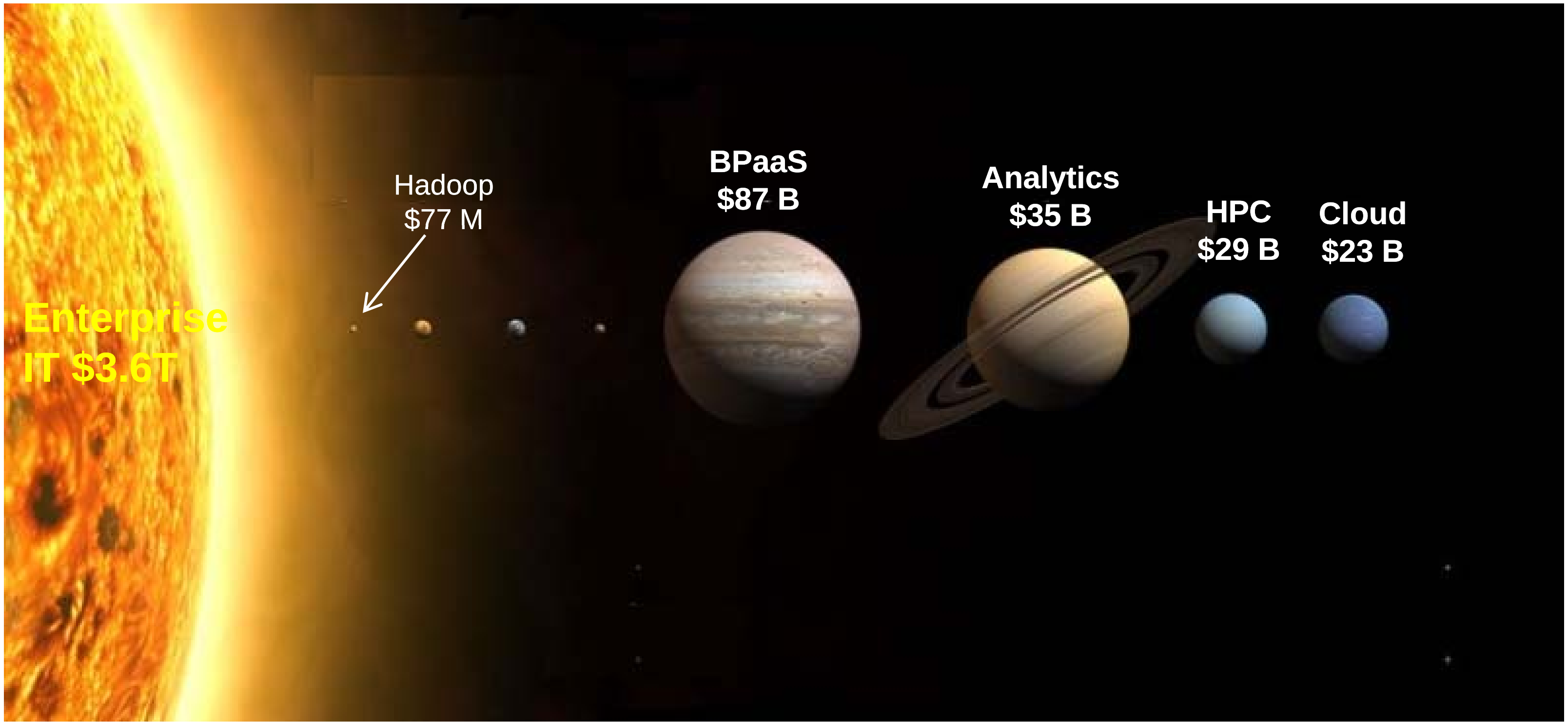
National Weather Service Data Stream

30 PB of New Data Annually

© 2012 NetApp, Inc. All rights reserved.

# Big Data = Big Analytics = Hadoop?

- That's What The Media Hype Implies, but it is NOT true!
- Traditional analytics (BI/DSS/DW) dominates the analytics market
- Like other technologies vying to gain broad adoption in Enterprise IT (e.g., Traditional Analytics, HPC & Cloud), it shows promise

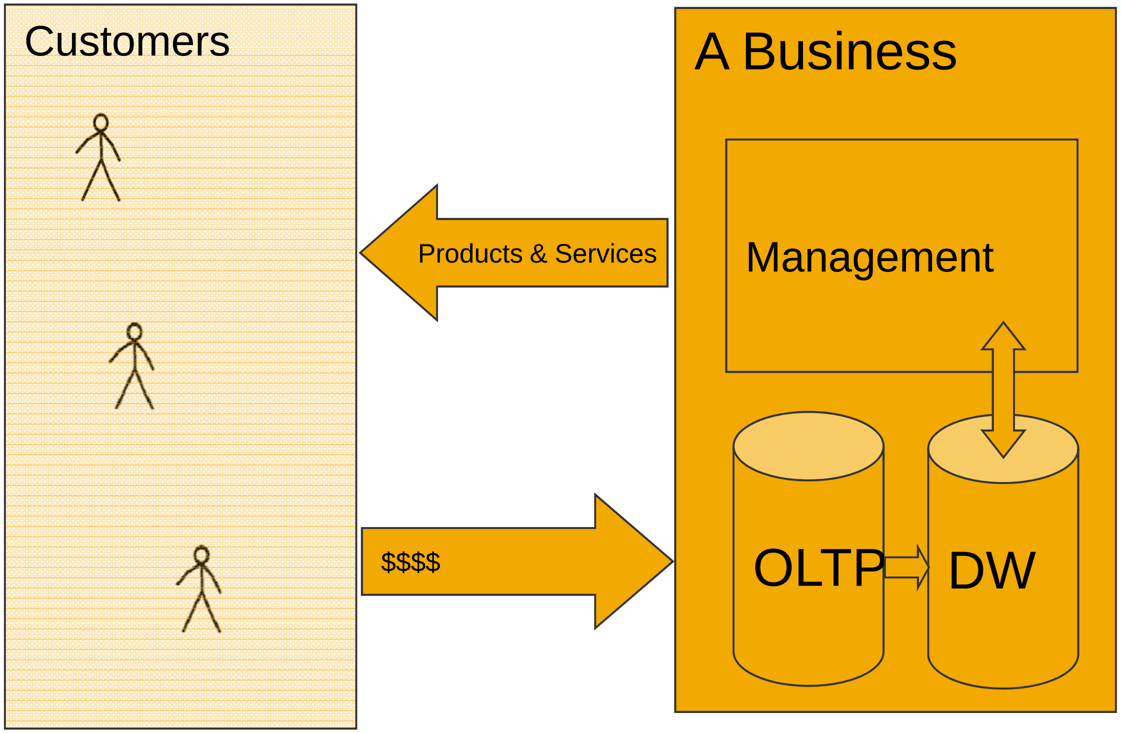


© 2012 NetApp, Inc. All rights reserved.

# Analytics

© 2012 NetApp, Inc. All rights reserved.

# Why Decision Support Systems are important?



DSS enables businesses to run "Closed Loop", ultimately improving their business through the use of feedback mechanisms.

© 2012 NetApp, Inc. All rights reserved.

NetApp

# Big Analytics – An Emerging Market

## NoSQL / Column DBs

- aster data — more data. big insights.
- Cassandra
- Couchbase
- HADAPT
- hadoop
- mongoDB
- ParAccel
- talend — open integration solutions
- VERTICA — An HP Company
- VoltDB

## Legacy DBs

- TERADATA
- ORACLE
- NETEZZA — an IBM Company
- GREENPLUM — A DIVISION OF EMC

## Middleware & Apps

- ATTIVIO — INSIGHT THAT MATTERS
- DIGITAL REASONING
- pentaho — open source business intelligence
- tableau
- INFORMATICA — The Data Integration Company
- QUEST SOFTWARE
- splunk>

## Open Source Distributors

- cloudera
- Hortonworks
- MAPR TECHNOLOGIES — EASY. DEPENDABLE. FAST.

## Cloud & Cyber

- amazon web services
- GOGRID
- Google
- KARMASPHERE
- KEYW
- tidemark
- YAHOO!

## Integration Services

- BAYSWATER — BUSINESS SOLUTIONS THAT EMPOWER
- Datameer
- lunexa — illuminate your data
- platfora
- ZALONI

## Compute

- hp
- DELL
- sgi
- StackIQ
- SUPERMICRO

## Network

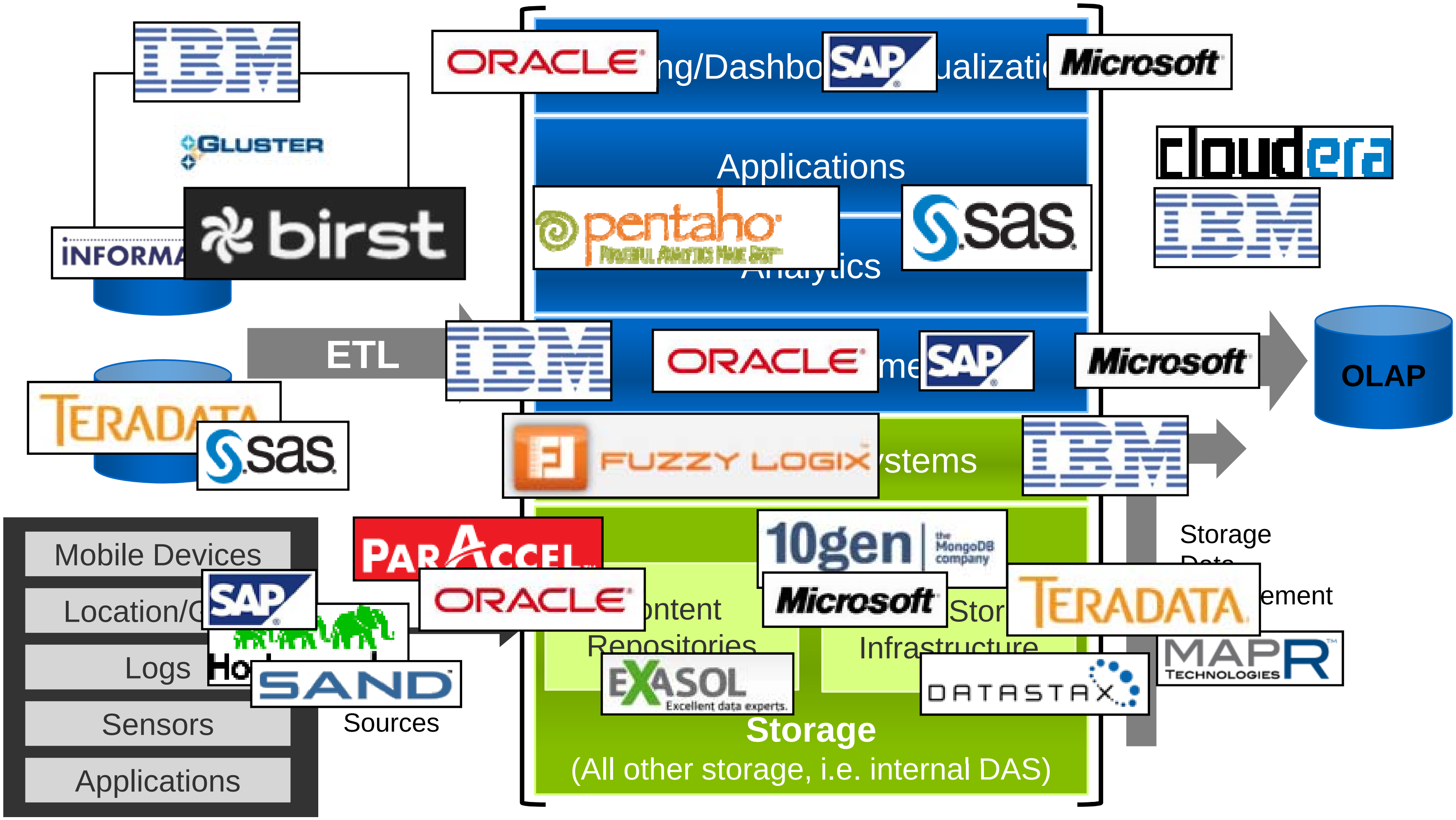
- CISCO
- ARISTA
- Mellanox TECHNOLOGIES
- SOLARFLARE

## Storage

- NetApp
- EMC²
- DataDirect NETWORKS

© 2012 NetApp, Inc. All rights reserved.

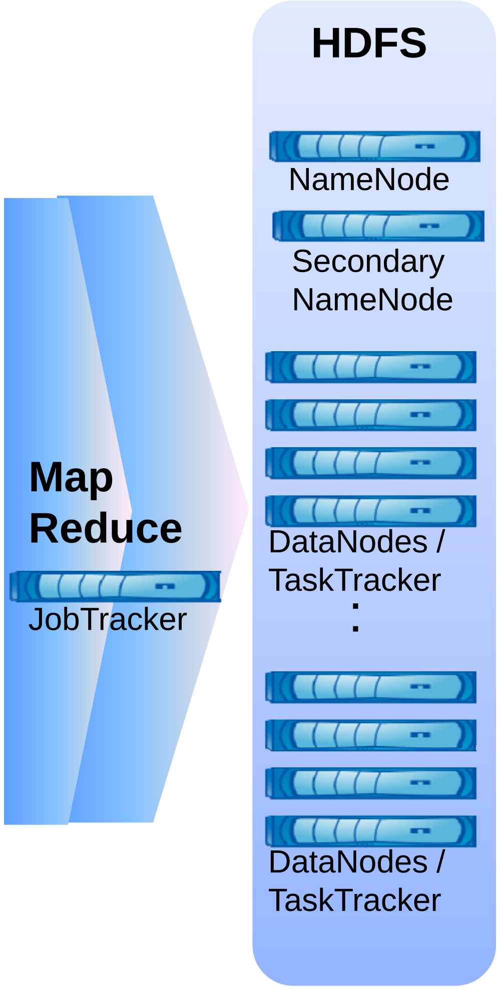
# Analytics & Enterprise Apps Environment



© 2012 NetApp, Inc. All rights reserved.

# What Does Hadoop Look Like Today?

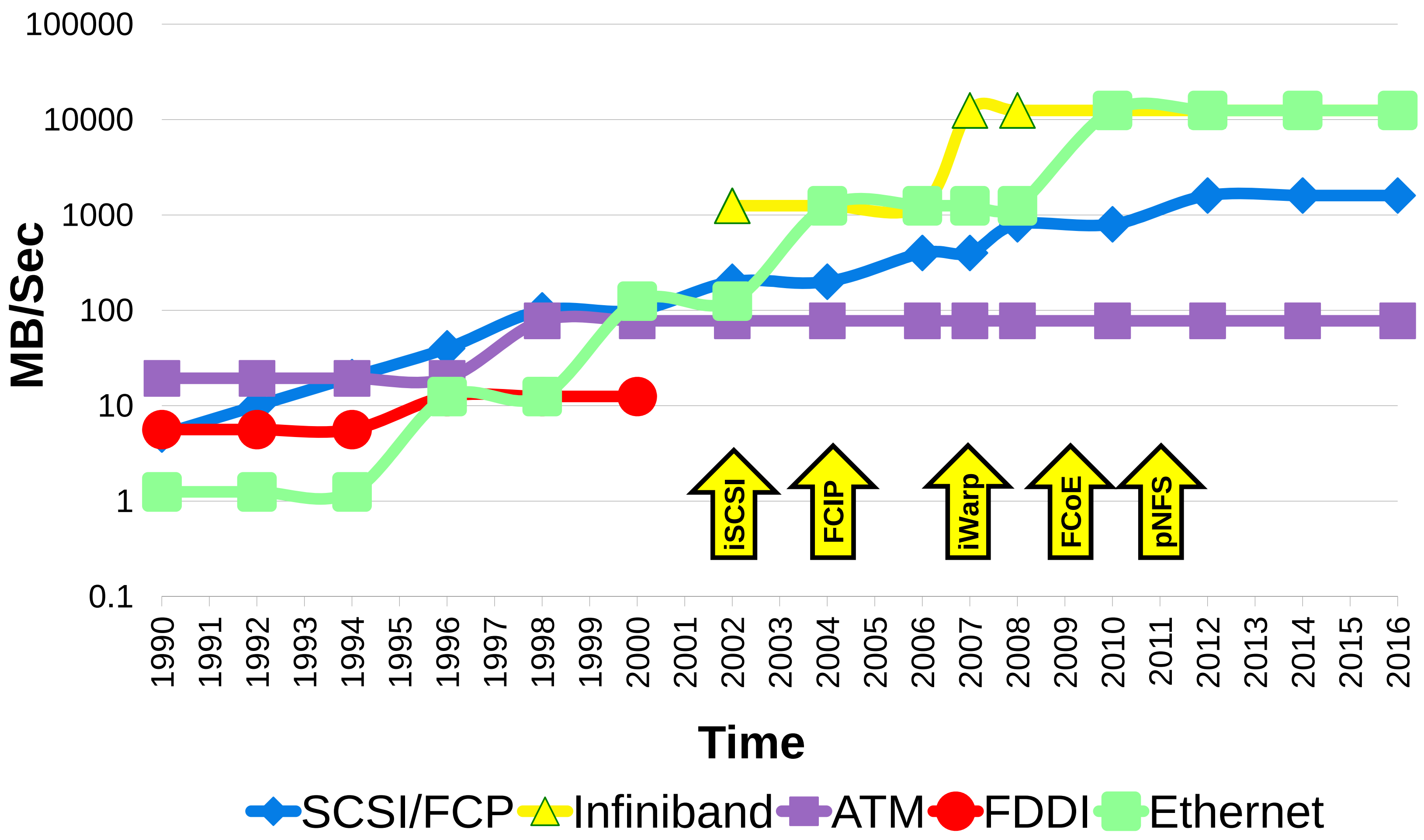
- Runs on a collection of cheap, commodity servers, in a distributed, shared nothing architecture
- Two key components
  - HDFS
    - Hadoop Distributed File System
  - MapReduce
    - Programming model for processing and generating large datasets



© 2012 NetApp, Inc. All rights reserved.

# Ethernet's Relentless March

Data will be growing by 50x, but bandwidth only by 10x!



© 2012 NetApp, Inc. All rights reserved.

# Why Should You Care?

## It’s the Value of your data

THOMSON REUTERS

5 Billion Records
Anywhere, Anytime
Faster time to market
50% Increase in Revenue

Sprint

Over 1PB of data
Growth of 175% YOY
90 days of data within
24 hours of a failure

- Top line revenue
  - Leverage their data assets into business advantage
- Bottom Line savings
  - Lower the cost of compliance
  - Manage ever growing data efficiently

© 2012 NetApp, Inc. All rights reserved.

# AutoSupport: Hadoop Use Case at NetApp

- “Call-home” service for all NetApp® systems
- Foundation of NetApp proactive support strategies
- Machine-generated data doubles every 16 months

| CHALLENGE | NETAPP SOLUTION | BENEFITS |
|---|---|---|
| 4 weeks to run a query on 24 billion unstructured records | <br>10-node Hadoop Cluster w/ shared Storage | Time reduced from 4 weeks to 10.5 hours |
| Impossible to run a query: 240 billion unstructured records | | Previously impossible, now achievable in just 18 hours |

“NetApp ASUP is a mission-critical application”

© 2012 NetApp, Inc. All rights reserved.

# Analytics of Tomorrow

- Traditional & Big Analytics side-by-side for years to come
- Hadoop moves to shared, virtualized infrastructure, for better efficiency and ease of management:
  - Hadoop remains logically distributed, shared nothing, but runs on a virtualized shared everything architecture (e.g., FlexPod for Vmware + eSeries)
  - Same as above, except Hadoop becomes logically shared everything, as HDFS is replaced by a parallel file system (e.g., Lustre Cluster, StorNext or GPFS)
- Enterprise class resiliency (no SPoF) and reliability with HPC-like performance (no need for triplicas)
- Use of a single copy of data for the map phase (higher storage utilization)
- Natural intersection with Cloud (Analytics as a Service)

© 2012 NetApp, Inc. All rights reserved.

# Summary

- Despite the hype, **Big Data is not new and is more than just analytics! (**Many agencies and private companies have struggled with Big Data for decades)
- **Analytics:** Traditional BI/DSS analytics still dominate. Importance of newer NoSQL & Columnar DB applications, enabled by MapReduce will grow with the growth of multi-structured data
- Big Data applications, such as Hadoop, will need to adopt shared, virtualized infrastructure (and its management benefits) if they are to be widely adopted by Enterprise IT

© 2012 NetApp, Inc. All rights reserved.

YOU’VE GOT

questions

?

I’VE GOT

rambling responses that sound like

answ

ers

@thebillp or william.peterson@netapp.com