IT 775
Database Technology

Data Stores

Big Data

WHAT IS BIG DATA?

Consider 1 day in your life

What is the best route to take?

What might the weather be?

What is the best way to invest money?

Should I take that loan?

Is there a way to do this task faster?

What have others done in similar cases?

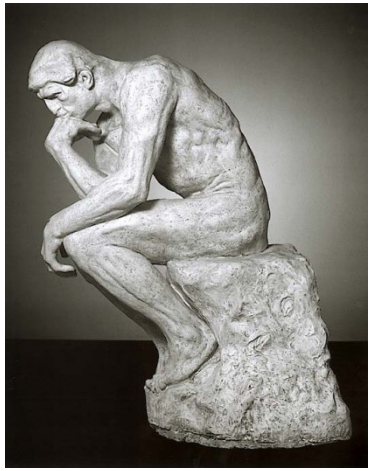
Which product should I buy?

Through the Ages, People Want…

To Know (what happened?)

To Explain (why it happened?)

To Predict (what will happen?)



Claim for Omniscience

To know, explain, and predict!

Grand challenge

Used other means to do this

Now, we try via science

Any sufficiently advanced technology is indistinguishable from magic.
---Arthur C. Clarke

We see a possibility through “lots of data”

Prophecies In Our Time

We can predict the weather

We can understand language translation pretty well

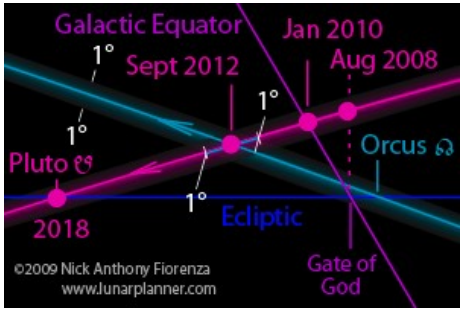
We can predict airflow well

enough to allow an airplane to fly

We can apply remedies to many common diseases

We can tell how astro bodies

(comets, planets) will behave



Weather

Remarkably accurate for a few days

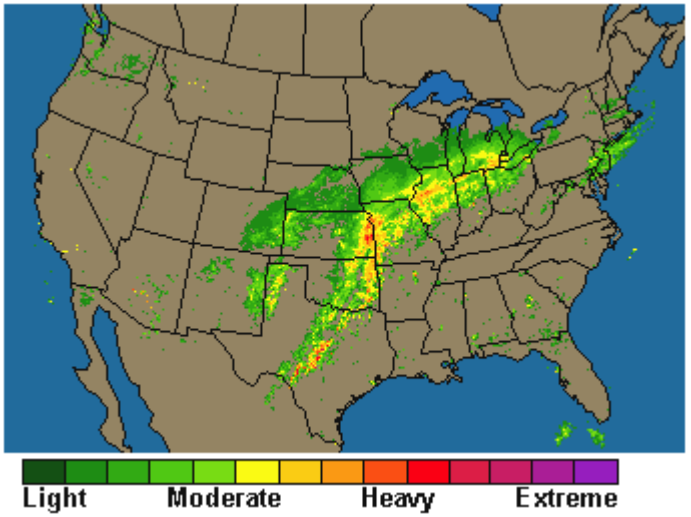
Data

Weather radar, satellite data,weather balloons, planes, etc.

Forecast Models

Simulation

Numerical methods



Computing Power is the Challenge

Algorithms take more than 24 hours to process

Resolution is the key

Democratizing Analysis

Forecasting was done, but in a limited manner in a few national laboratories and others

That is changing!

Big Data

Everyone seems to be talking about it, but what is big data really?

How is it changing the way researchers at companies, non-profits, governments, institutions, and other organizations are learning about the world around them?

Where is this data coming from, how is it being processed, and how are the results being used?

And why is open source so important to answering these questions?

What is Big Data?

No rule saying how “big” a database needs to be to be “big”

Requires new techniques and tools to process data

Need programs that span multiple physical/virtual machines working together in concert to process all of the data in a reasonable span of time

Getting programs on multiple machines to work together in an efficient way, so that each program knows which components of the data to process, and then being able to put the results from all of the machines together to make sense of a large pool of data takes special programming techniques.

Since it is typically much faster for programs to access data stored locally instead of over a network, the distribution of data across a cluster and how those machines are networked together are also important considerations which must be made when thinking about big data problems.

Data is the Wealth of Our Time

"Data is a precious thing because they last longer than systems"

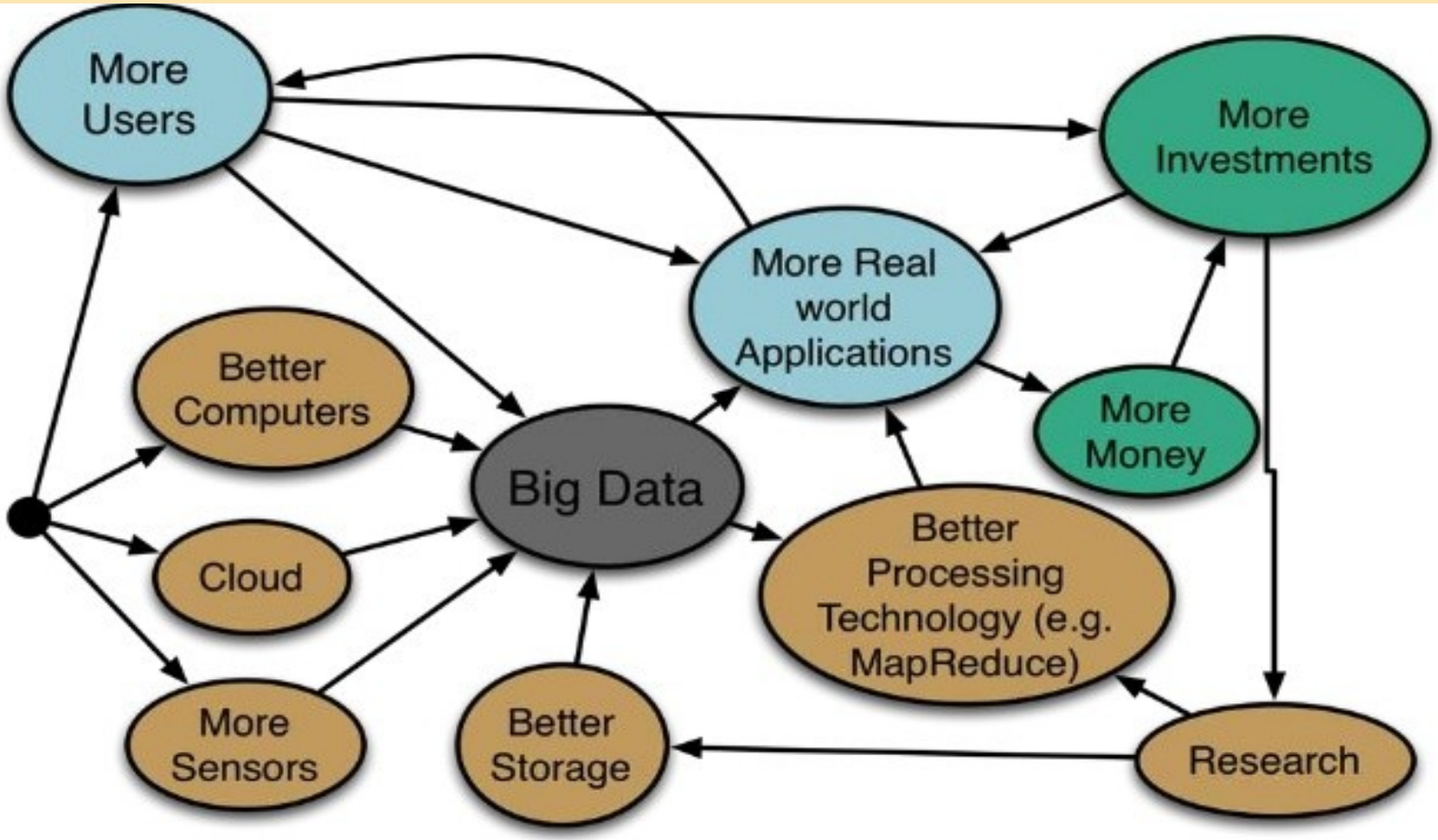
---Tim Berners-Lee

Access to data is becoming ultimate competitive advantage

Google+ versus Facebook

Why many organizations try hard to give us free things and keep us always logged in (e.g. Gmail, Facebook, search engine toolbars, Coke, McDonalds)

Drivers of Big Data



Data Avalanche

We are now collecting and converting large amounts of data to digital form

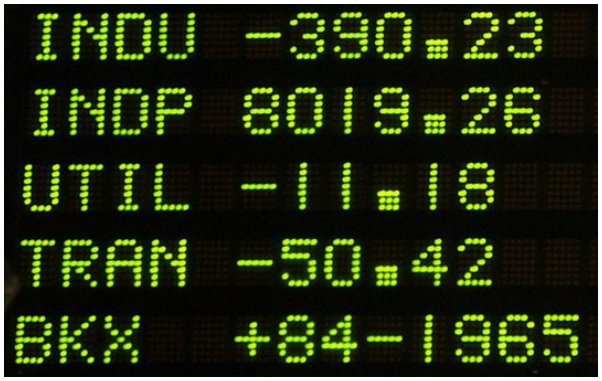
90% of the data in the world today was created within the past 2 years (18.4 Zettabytes (1e21) in 2017)

Amount of data doubles every 2 years

In Real Life, Most Data are Big

- Web does millions of activities per second, filling server logs
- Four-engine jumbo jet, crossing the Atlantic, creates 640 terabytes of data. With >25,000 flights daily
- Social network users
- There are >4 billion cell phones
- Billions of RFID tags

Datasets



But the potential uses go much further!

Transactional Data
stock prices
bank data
individual merchants' purchase histories

Sensor Data
the Internet of Things (IoT)
measurements from manufacturing line robots of auto maker
location data on a cell phone network
instantaneous electrical usage in homes and businesses (smart meter)
passenger boarding information taken on a transit system

By analyzing this data, organizations are able to learn trends about the data they are measuring, as well as the people generating this data. The hope for this big data analysis are to provide more customized service and increased efficiencies in whatever industry the data is collected from.

How do we Analyze Big Data?

MapReduce

a method for taking a large data set and performing computations on it across multiple computers, in parallel

a model often used to refer to the actual implementation of this method

MapReduce consists of two parts

Map function

- sorting and filtering
- placing data in categories for analysis

Reduce function

- summary of the data; combining it all together

While largely credited to research which took place at Google, MapReduce is now a generic term and refers to a general model used by many technologies.

What About Moving Data?

We want to process data as it is received in a streaming fashion

Very fast ouput

Lots of events (100k to millions)

Processing without storage

Two main technologies

Stream processing (http://storm.apache.org/)

Complex Event Processing (CEP) http://wso2.com/products/complex-event-processor/

Big Data Challenges

Speed

Extracting semantics and handling multiple representations and formats

Security data ownership, delegation, permissions and privacy

Making data accessible to all parties anywhere, anytime, from any device through any format

Is MapReduce good enough? What about other parallel problems?

Handling uncertainty

Why is Big Data so hard?

How to store? Assuming 1TB, it takes 1000 computers to store 1PB

How to move? Assuming 10Gb network, it takes 2 hours to copy 1TB or 83 days to copy 1PB

How to search? Assume each record in db is 1KB and 1 computer can process 1000 records per second, we need 277 CPU days to process 1TB and 785 CPU years to process 1PB

How to process?

How to convert serial algorithms to work in large size and in parallel

How to create new algorithms

Big Data Architecture