

Big Data Management

After this video you will be able to..

- Describe what “data management” means
- Identify the primary issues involved in the management of “big data”

What is Data Management?

Must-Ask Questions about a Data Application

How do we **ingest** the data?

Where and how do we **store** it?

How can we ensure **data quality**?

What **operations** do we perform on the data?

How can these operations be **efficient**?

How to **scale up** data volume,
variety, velocity and access?

How to keep the data **secure**?

Ingestion Infrastructure

How many data sources?

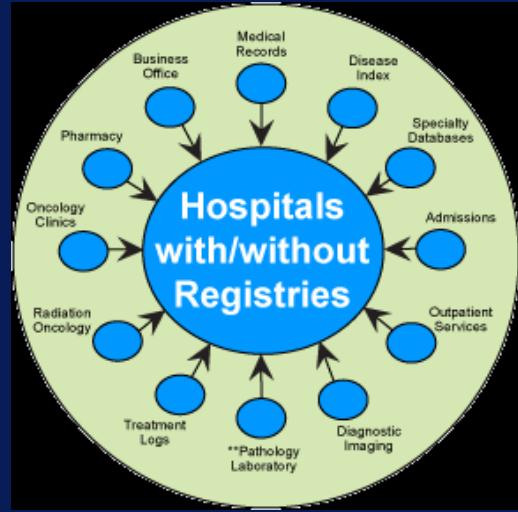
How large are data items?

Will the number of data sources grow?

Rate of data ingestion?

What to do with bad data?

What to do when data is too little or too much?



An imaginary cloud database of personal information

Ingestion Infrastructure

How many data sources? ~20

How large are data items?

Avg. record size: 5KB, Avg. image size:
2GB, #records: 50 Million

Will the number of data sources
grow? Not much

Rate of data ingestion? 3k/day

What to do with bad data?
Warn, flag and ingest

What to do when data is too little or
too much?
Not likely

Ingestion policy



Ingestion Infrastructure

How many data sources? 2M

How large are data items?

Avg. record size: 3KB, Avg. image size:
2MB, #records: 200 Billion

Will the number of data sources
grow?

Now 25 M, growing at 15% per year

Rate of data ingestion? peak 200k/hr

What to do with bad data?
Retry once, then discard

What to do when data is too little or too
much?

Spill to auxiliary server for 10 TB, reclaim
lazily, drop by 0.1% steps when > 85% full



An imaginary cloud database of personal
information

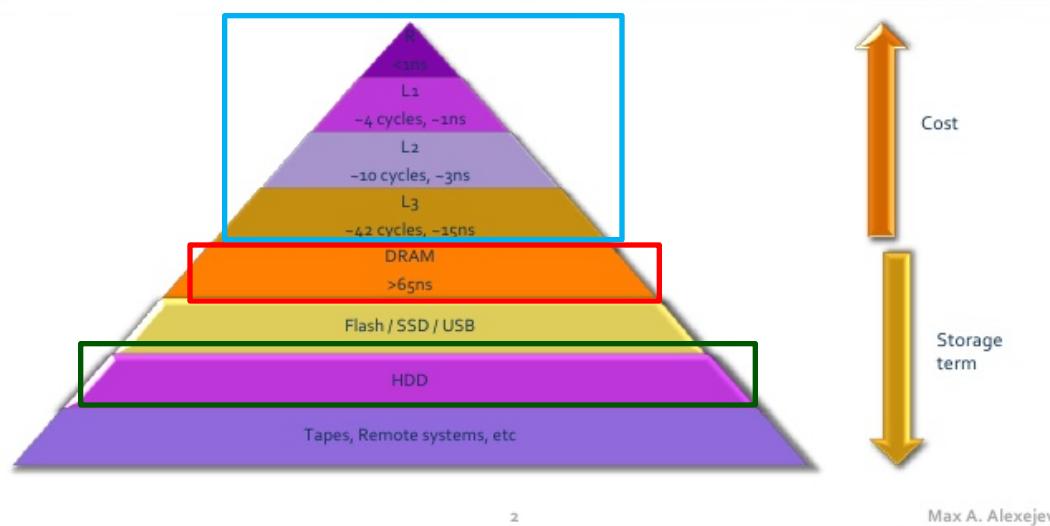
Storage Infrastructure

where hardware meets Data Management

How much data to store?

Directly attached? Network attached?

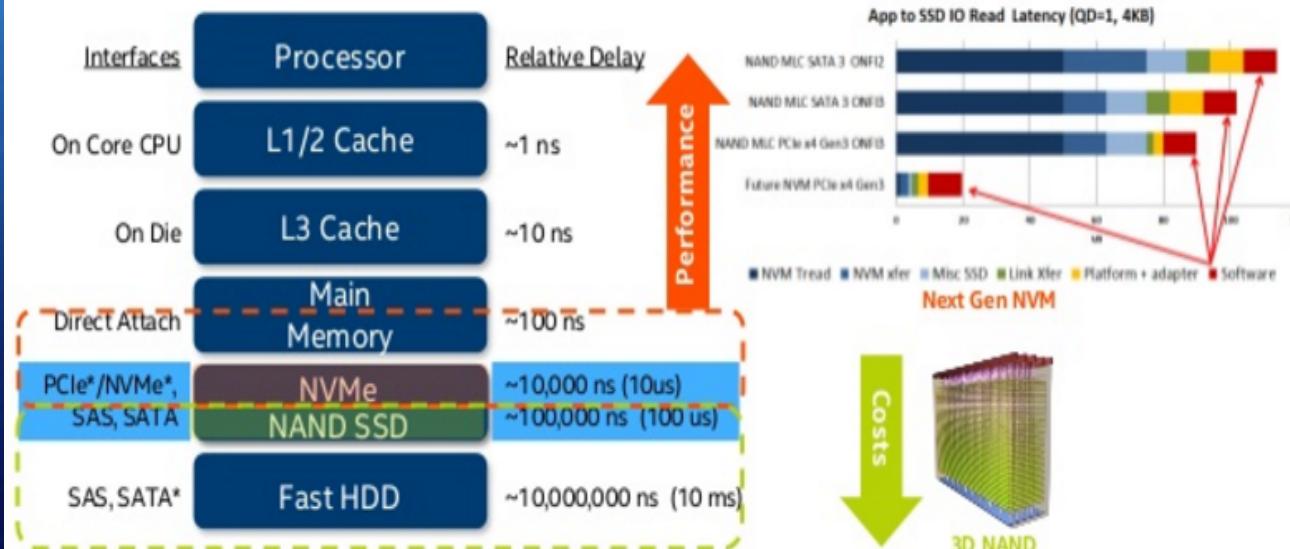
Memory Hierarchy



SSD: Solid State Device

How fast do we need to read/write?

Future Memory and Storage Hierarchy



*NVMe: Non-volatile Memory Express
For fast transfer between memory and SSD*

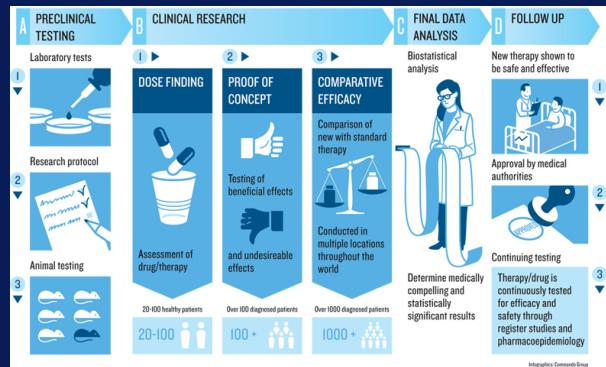
Data Quality

Why worry about data quality?

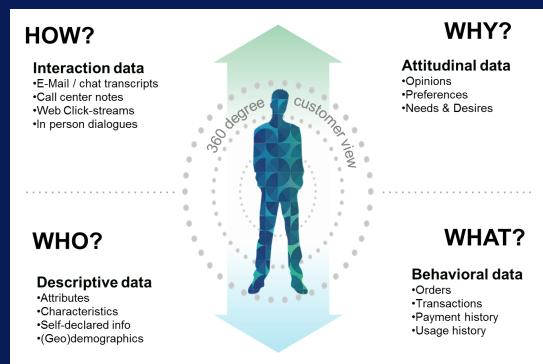
Better quality means better analytics and decision making



Quality assurance means needed for regulatory compliance



Quality leads to better engagement and interaction with external entities



- **Data profiling and data quality measurement:** The analysis of data to capture statistics (metadata) that provide insight into the quality of data and help to identify data quality issues.
- **Parsing and standardization:** The decomposition of text fields into component parts and the formatting of values into consistent layouts, based on industry standards, local standards (for example, postal authority standards for address data), user-defined business rules, and knowledge bases of values and patterns.
- **Generalized "cleansing":** The modification of data values to meet domain restrictions, integrity constraints or other business rules that define when the quality of data is sufficient for an organization.
- **Matching:** The identifying, linking or merging of related entries within or across sets of data.
- **Monitoring:** The deployment of controls to ensure that data continues to conform to business rules that define data quality for an organization.
- **Issue resolution and workflow:** The identification, quarantining, escalation and resolution of data quality issues through processes and interfaces that enable collaboration with key roles, such as data steward.
- **Enrichment:** The enhancement of the value of internally held data by appending related attributes from external sources (for example, consumer demographic attributes and geographic descriptors).

Operations on Data

Operations on single data items
that produce a sub-item



subarray

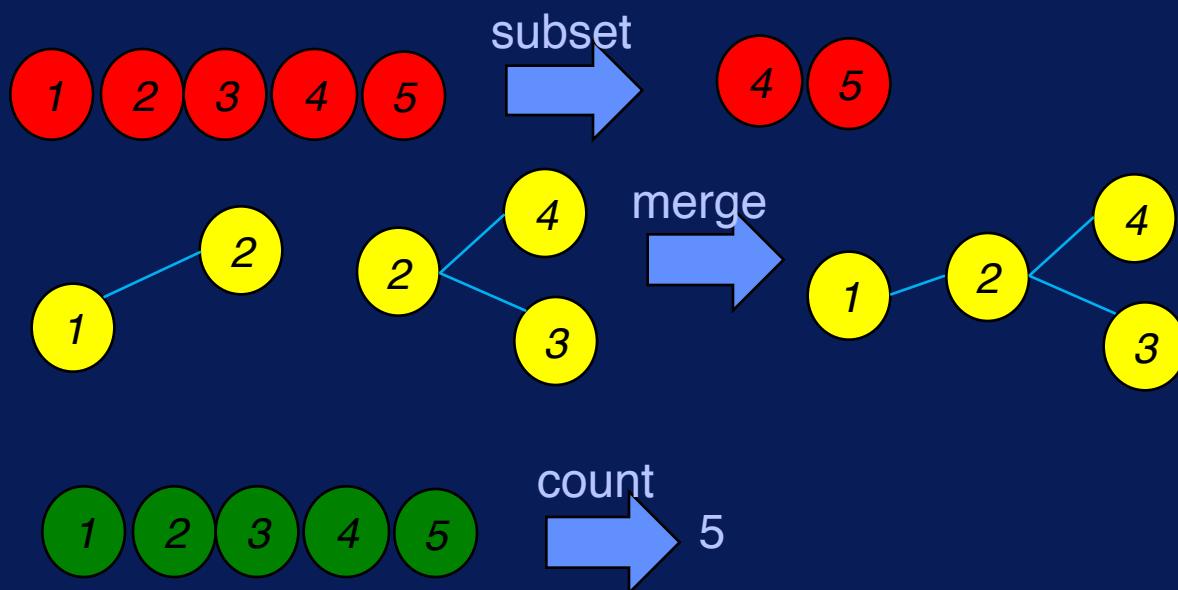


Operations on collections of
data items

Operations that select a part of
a collection

Operations that combine two
collections

Operations that compute a
function on a collection

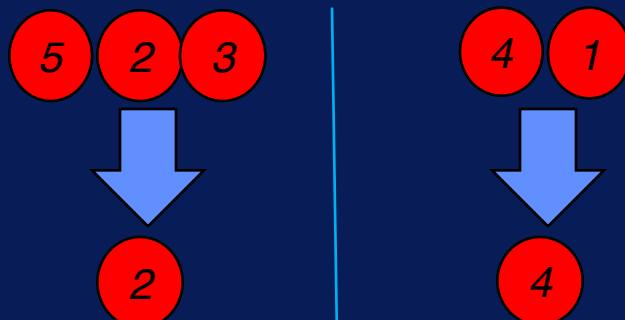


Efficiency of Data Operations

Measured by time and space

Should use parallelism

Selection



Achieving Scalability

Vertical Scaling (Scale-up): Adding more processors and RAM, buying a more expensive and robust server

Many operations perform better with more memory, more cores

Maintenance can be difficult, expensive

The Server industry has many solutions for scale-up/scale-out decisions

Scaling up and Scaling Out

Horizontal Scaling (Scale-out): Adding more, possibly less powerful machines that interconnect over a network

Parallel operations will possibly be slower

Easier in practice to add more machines

Keeping Data Secure

Data security – a must for sensitive data

Increasing the number of machines leads
to more security risks

Data in transit must be secure

Encryption and decryption increase
security but make data operations
expensive