www.hops.io

@hopshadoop

Genomics for Big Data and Hops Hadoop

Jim Dowling
Associate Prof @ KTH
Senior Researcher @ SICS RISE

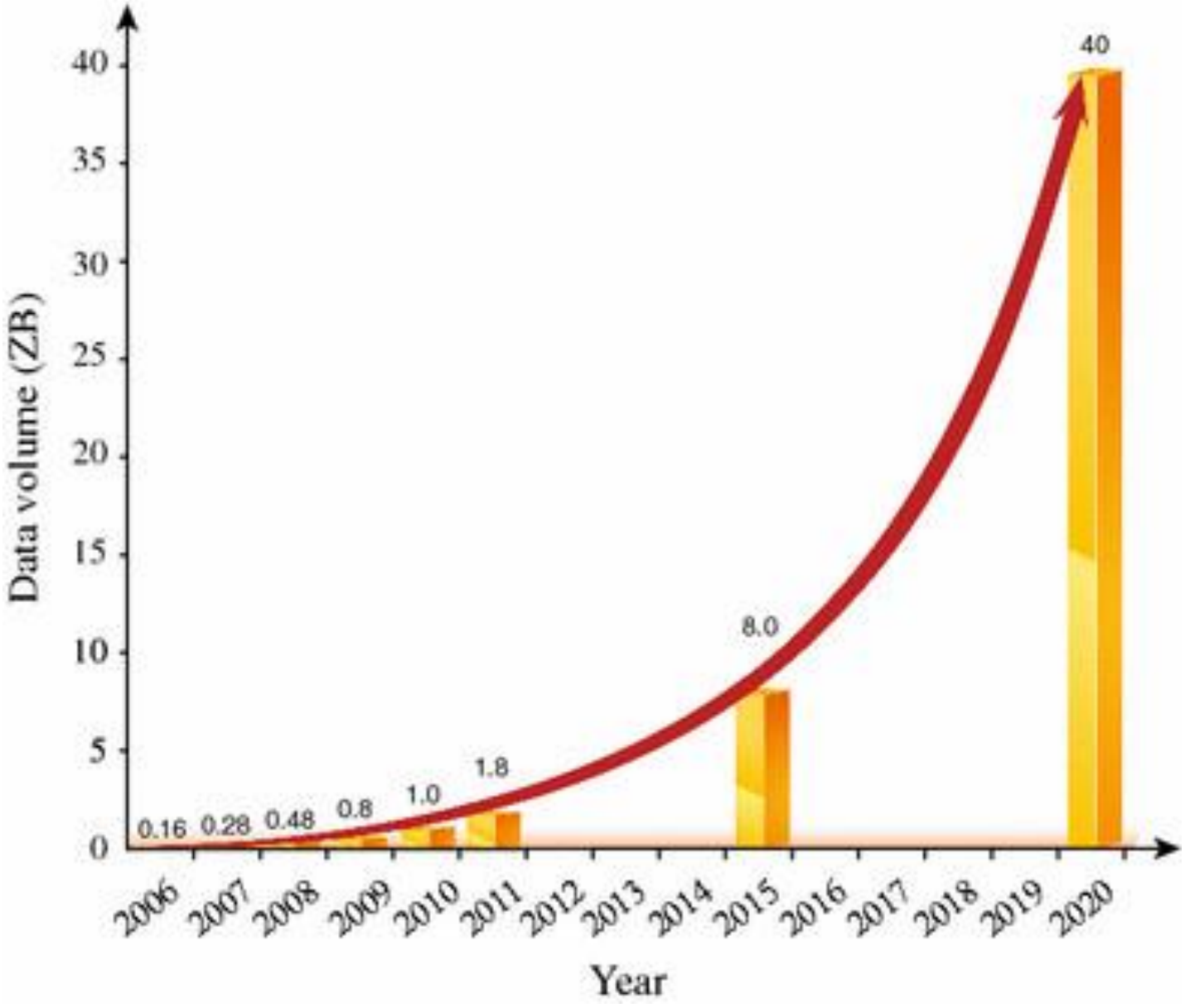
EMBnet COST.CHARME training school
“Big Data for Life Sciences”

Hops Hadoop

- World-record performance in Hadoop*
  - 16-37X throughput of the Hadoop Filesystem (HDFS)
  - Exabyte size Clusters
- Self-Service
  - Based on new concepts: Projects, Datasets, Project-Users
- Running in Production > 1 Year
- Startup commercializing Hops, Logical Clocks AB

*https://www.usenix.org/conference/fast17/technical-sessions/presentation/niazi

Growth in Data Volumes



Why is Big Data Important?

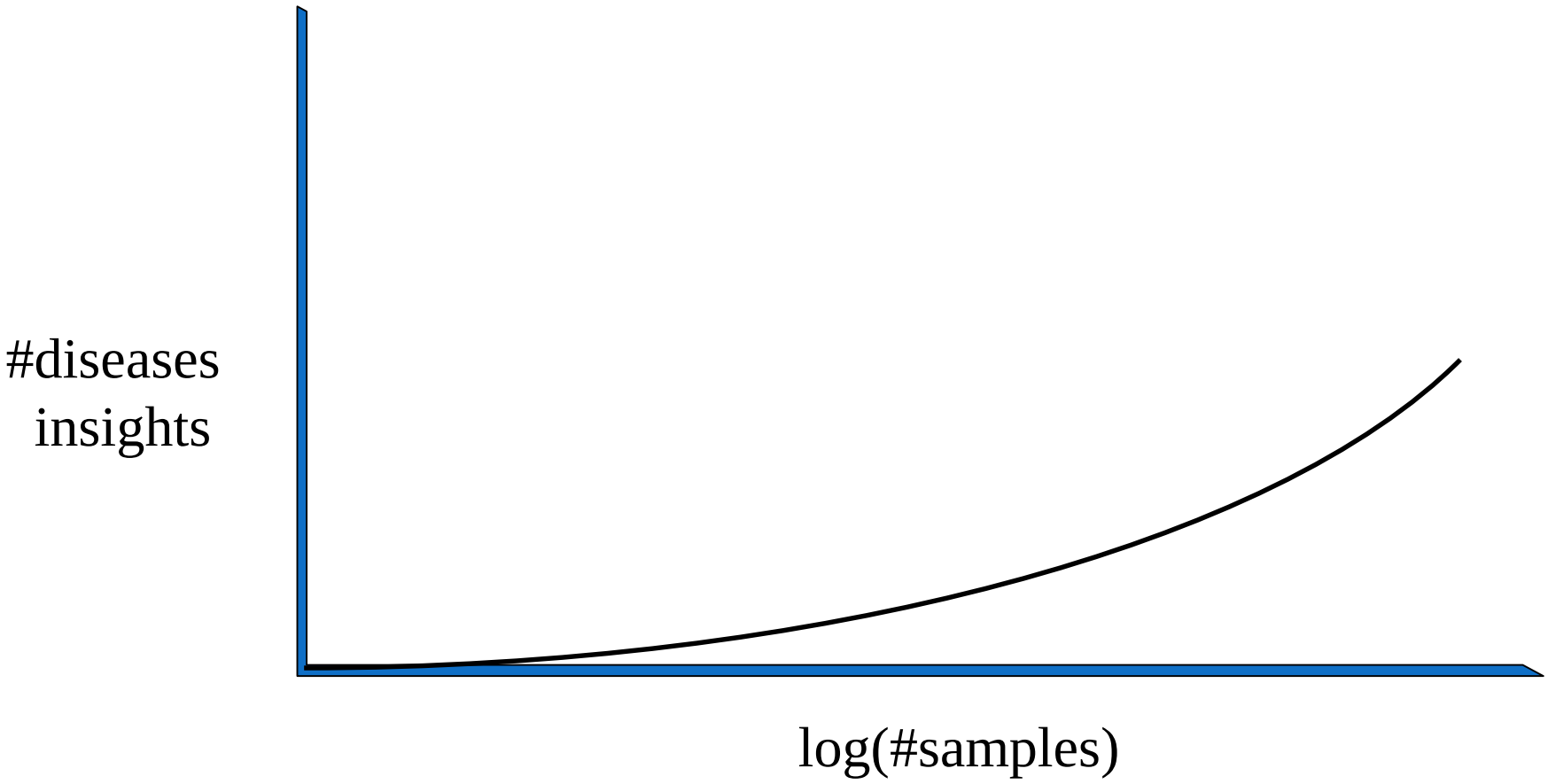
- In a wide array of academic fields, the ability to effectively process data is superseding other more classical modes of research.

“More data trumps better algorithms”*

*“The Unreasonable Effectiveness of Data” [Halevey, Norvig et al 09]

**“Revisiting the Unreasonable Effectiveness of Data” [2017]

Bigger Datasets have more Statistical Power



Big Data Explosion!

Lots of potential

But…..

Output needs to be stored for many years

Risks for those who store it

One major incident away from scaring the public off entirely

*Analogy from Maciej Ceglowski (Pinboard.in)

$1000 per Whole Human Genome

HiSeq X Ten^ => ~18,000 genomes/year

Volume => ~5.2 PB/year*

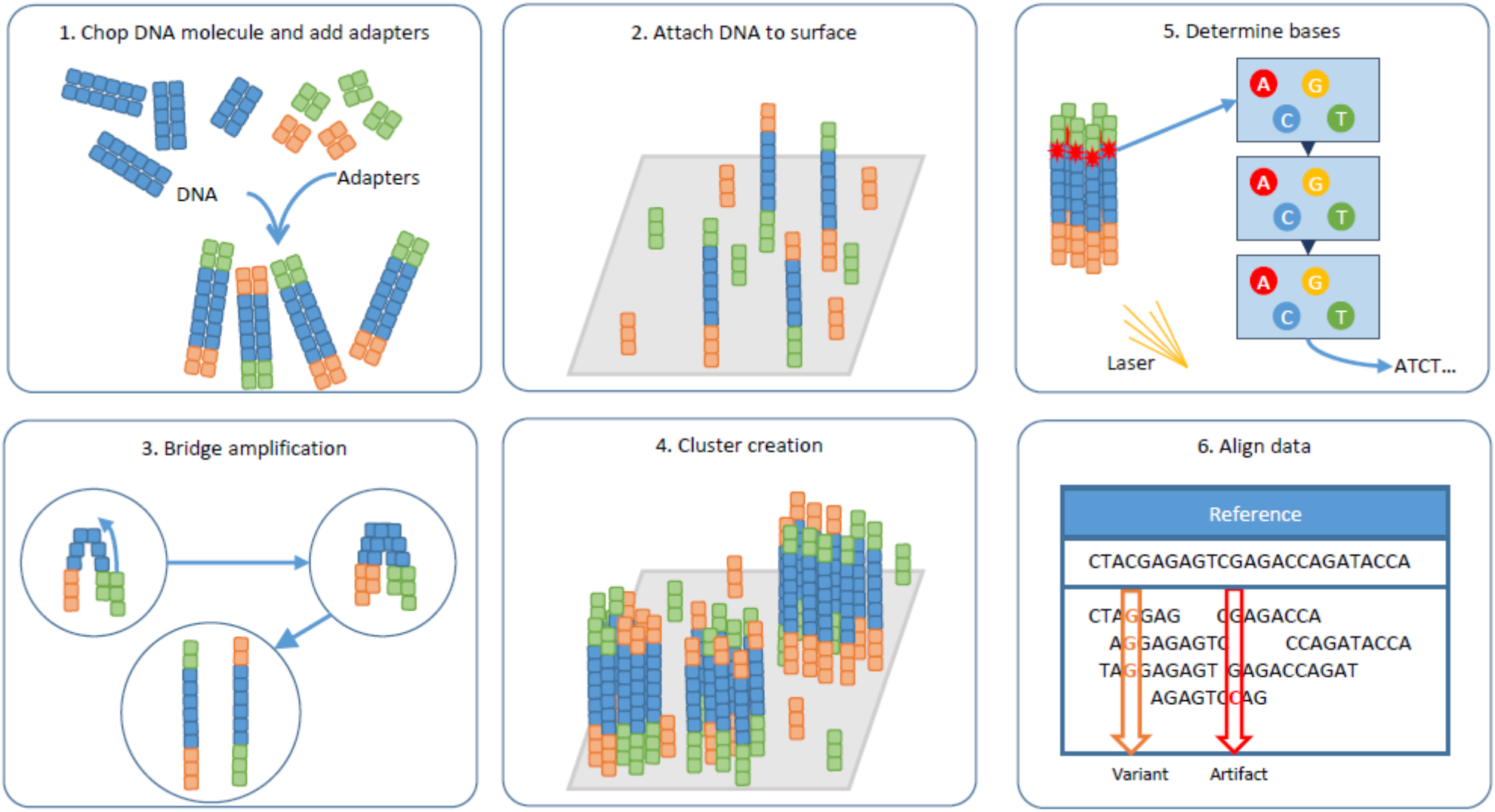
Velocity => ~45 MB/sec*

^Cost ~$10 million

*5.2 PB assumes a replication factor of 3

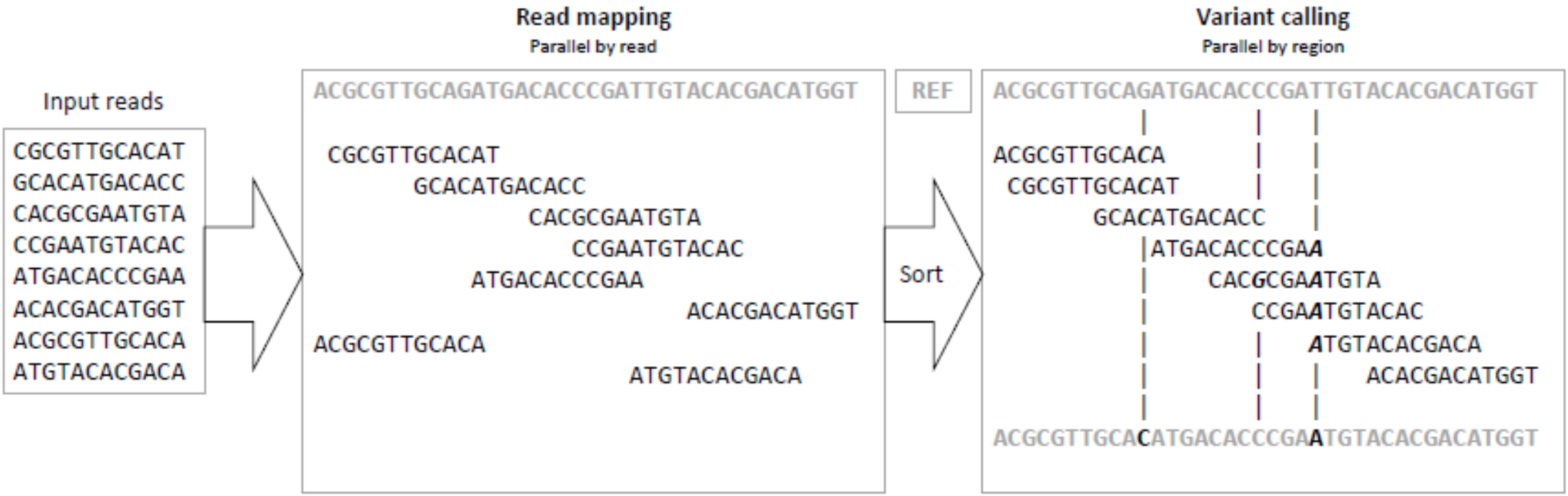
See: http://goo.gl/OCgJ36

Illumina Sequencing Steps



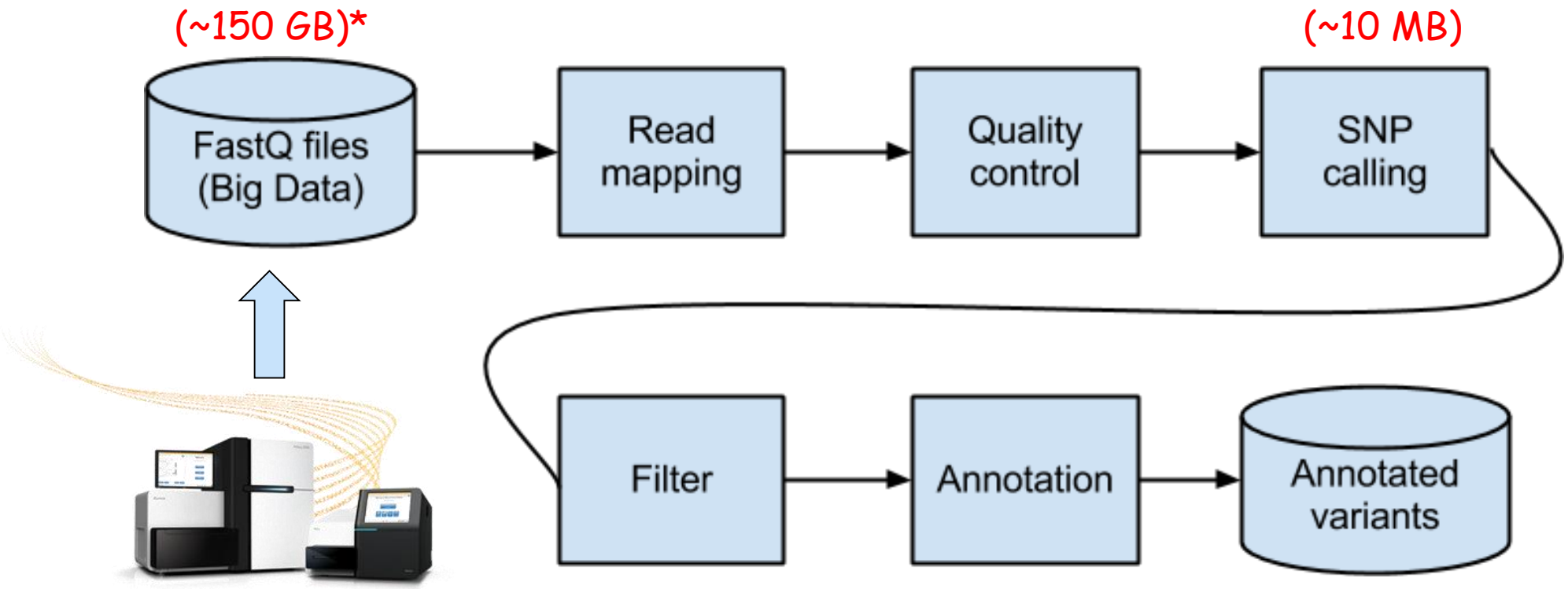
[Figure 1, Phd Thesis, Dries Decap, Univ Ghent 2017]

Whole Genome Sequencing Processing Steps



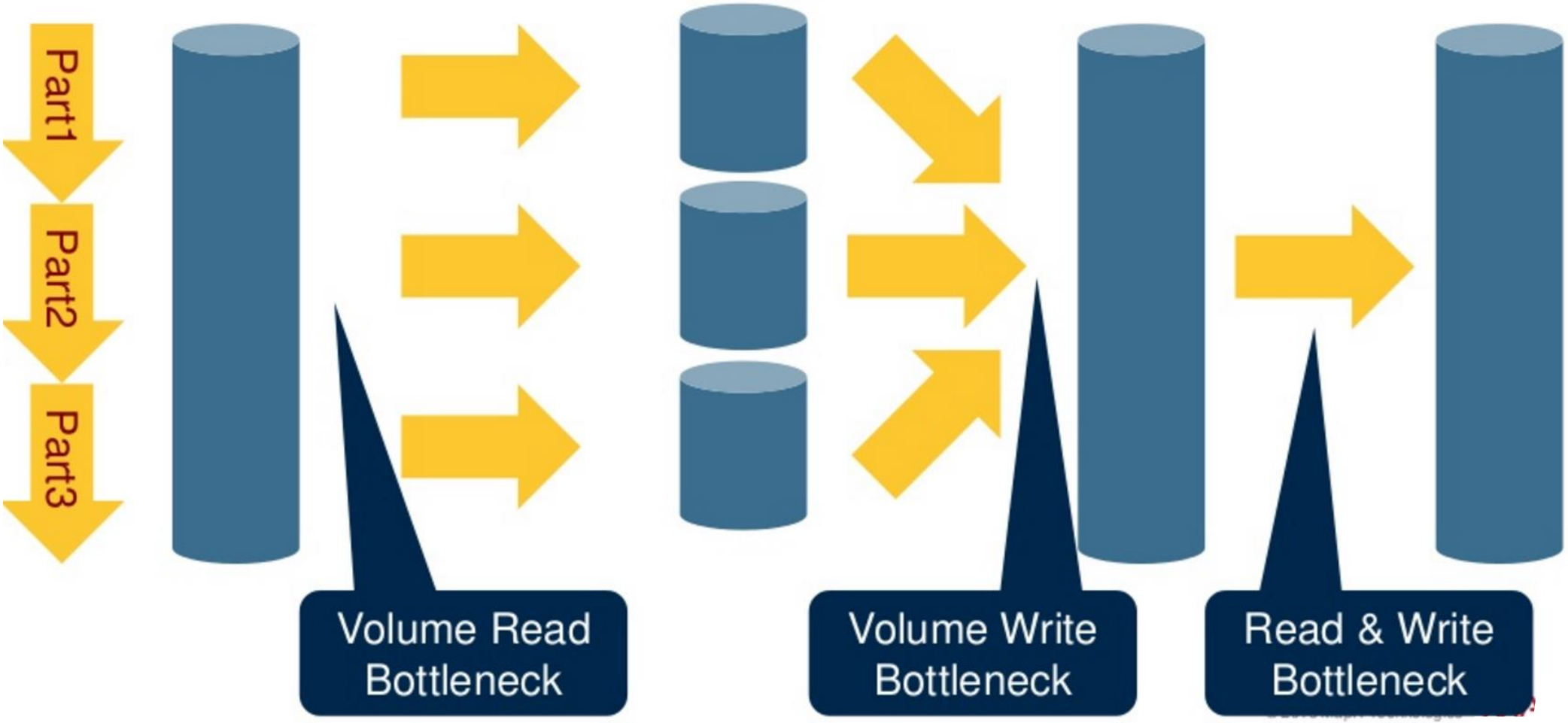
[Figure 2, Phd Thesis, Dries Decap, Univ Ghent 2017]

Whole Genome Sequencing: The Petabyte Era



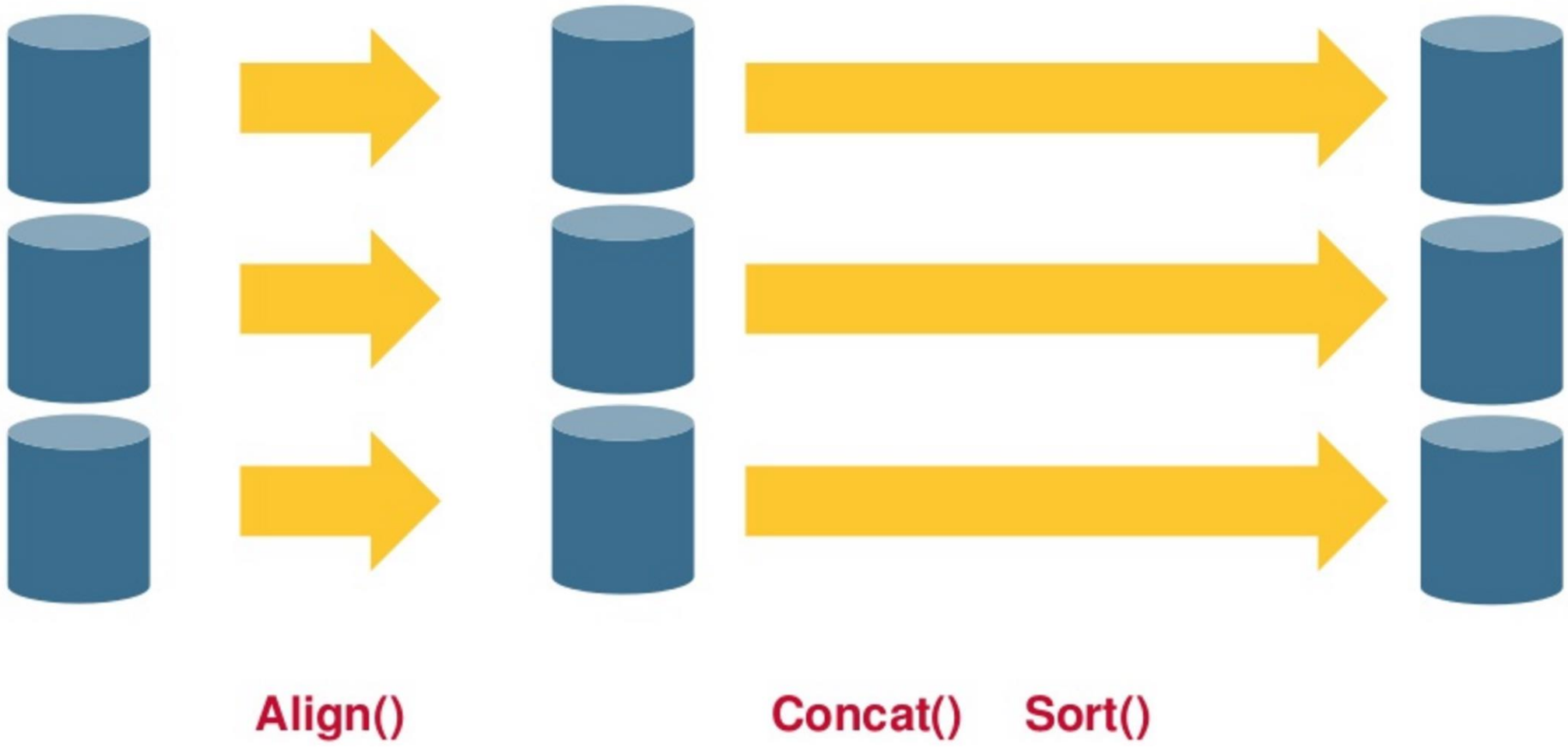
*Assuming 75X coverage for WGS

Whole Genome Sequencing Pipeline



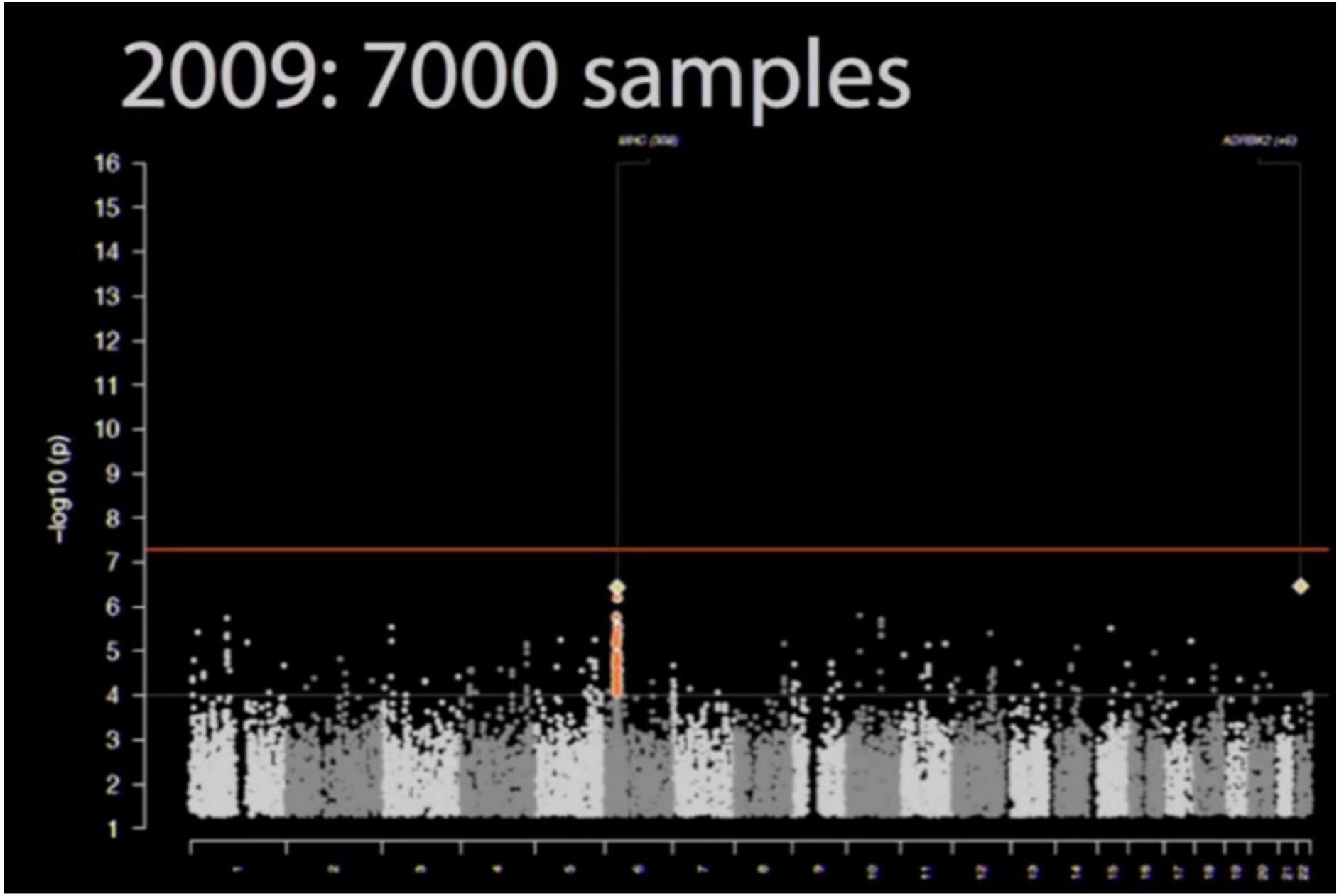
[From MapR]

Whole Genome Sequencing Pipeline



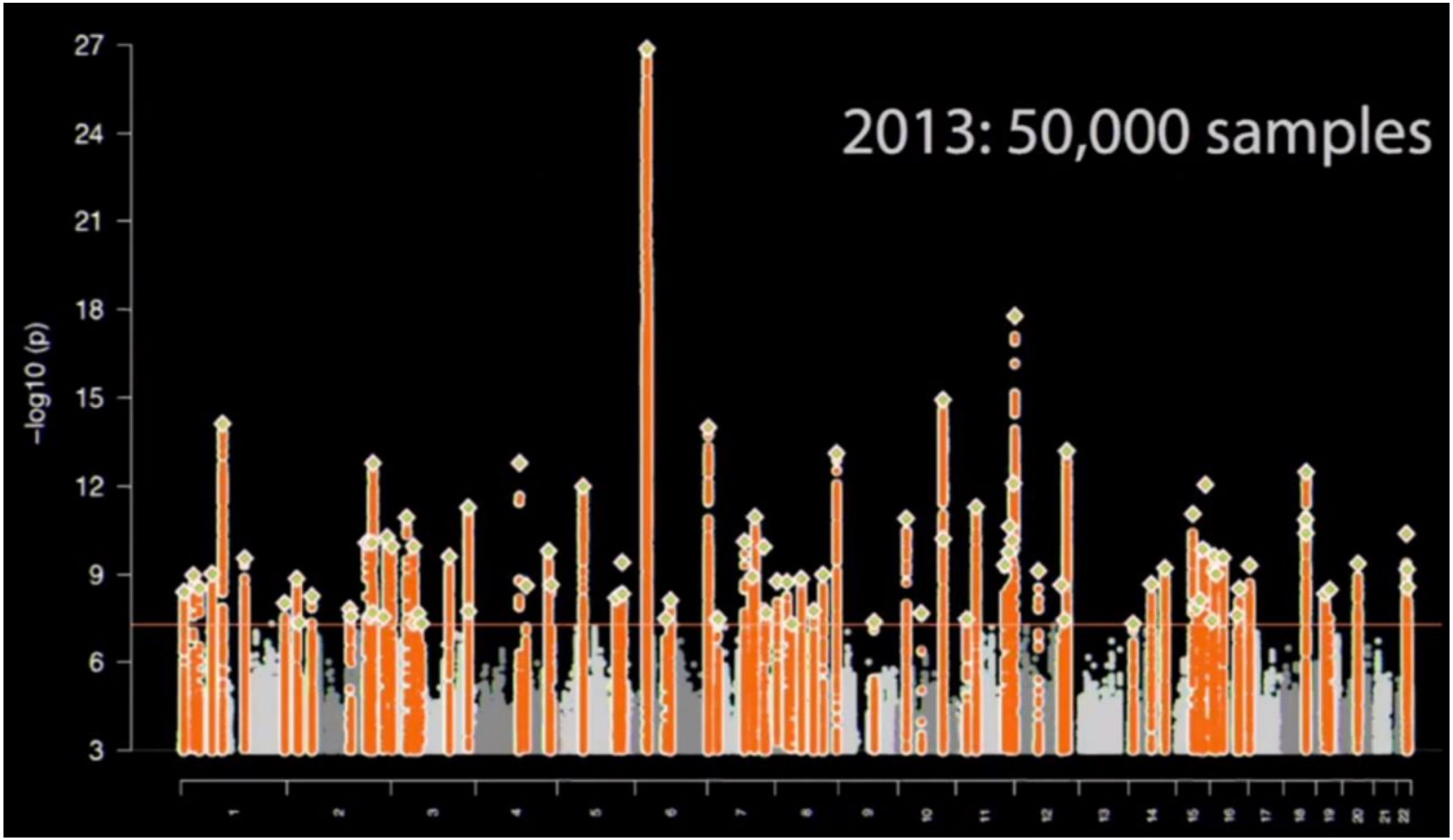
[From MapR]

Genomics needs Big Data



[Image source: Patterson, Fighting the Big C with the Big D, 2014]

Genomics needs Big Data

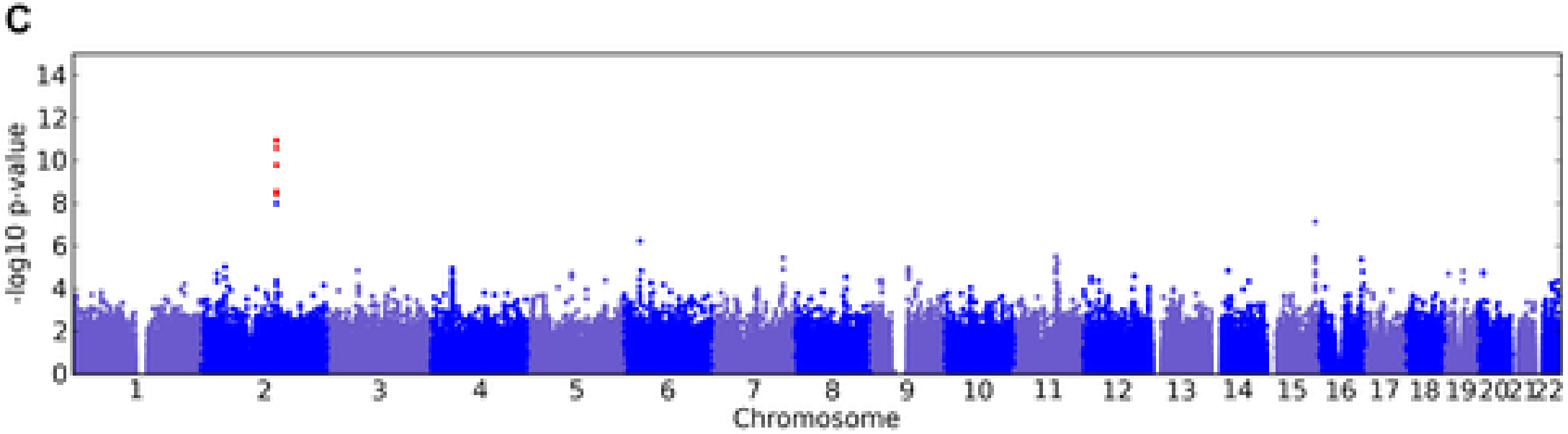


[Image source: Patterson, Fighting the Big C with the Big D, 2014]

Bigger Datasets as Hypothesis Generators

Photic sneeze reflex SNPs discovered by 23andme*

- rs10427255, near *ZEB2*, and rs11856995, near *NR2F2*



*http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000993

Big Data is Commodity Hardware

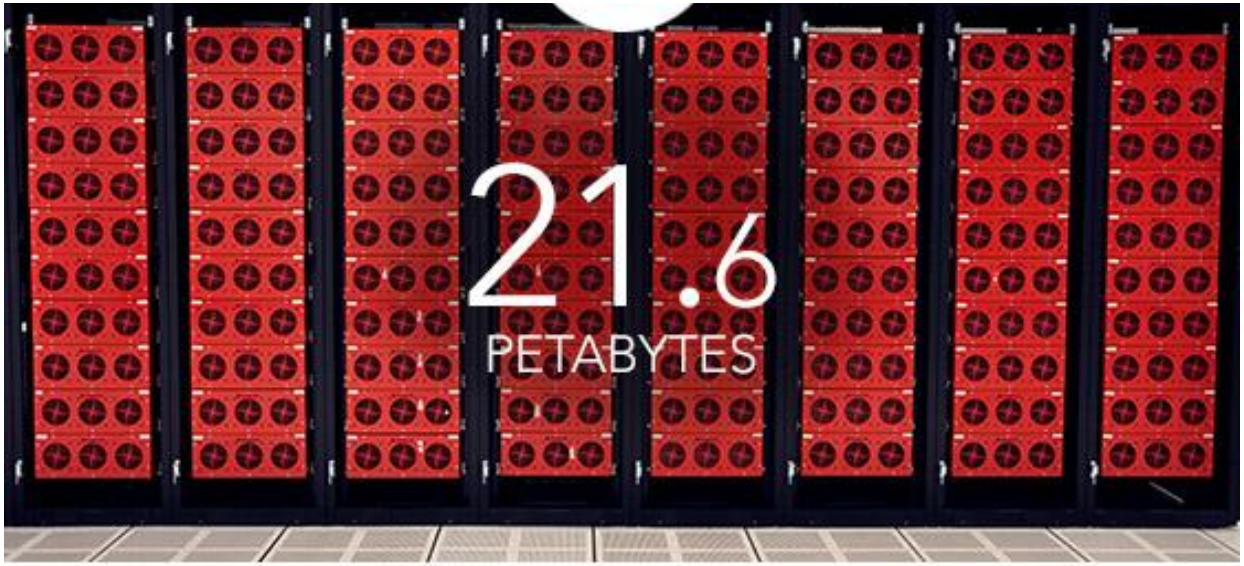
180TB for $9,305

http://blog.backblaze.com/2014/03/19/backblaze-storage-pod-4

5PB for $279,150

http://blog.backblaze.com/2014/03/19/backblaze-storage-pod-4

~133K Whole Genomes*: $1M (21.6PB)



*Each genome is 112.5 GB, 35x coverage, replicated with Reed-Solomon erasure-coded.

GPUs are going Commodity as we speak..

- Commodity Server*
  - 10 Nvidia GTX 1080Ti
    - 11 GB Memory
  - 256 GB Ram
  - 2 Intel Xeon CPUs
  - 56 Gb/s Mellanox
  - SingleRoot PCI Complex

- Nvidia DGX-1
  - 8 Nvidia Tesla P100/V100
    - 16 GB Memory
  - 512 GB Ram
  - 2 Intel Xeon CPUs
  - 56 Gb/s Mellanox
  - NVLink

10 x Commodity Server = 150K Euro

Price per DGX-1 = 150K Euro

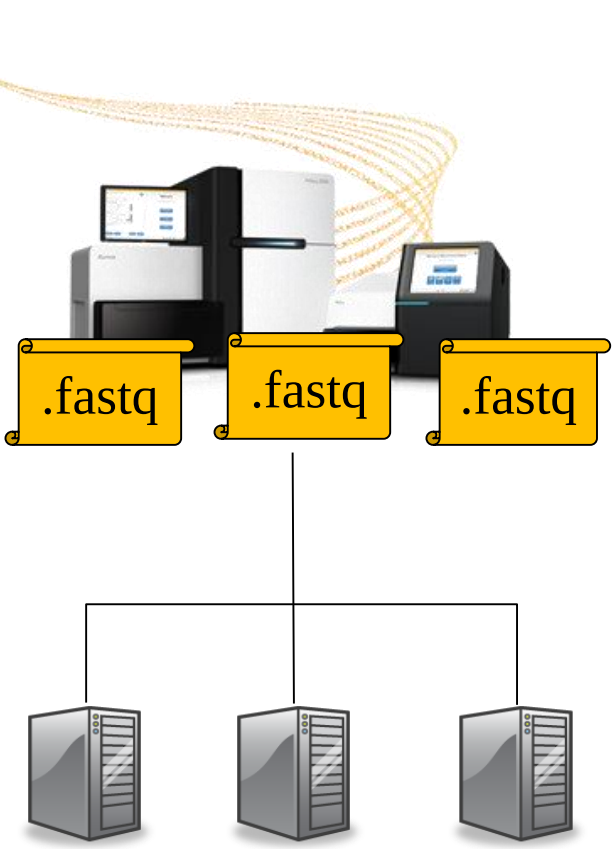
*https://www.servethehome.com/single-root-or-dual-root-for-deep-learning-gpu-to-gpu-systems/

Big Data Administration Costs

Facebook Operations staffers manage 20-26,000 servers each^

^ http://allfacebook.com/20000-servers_b127053

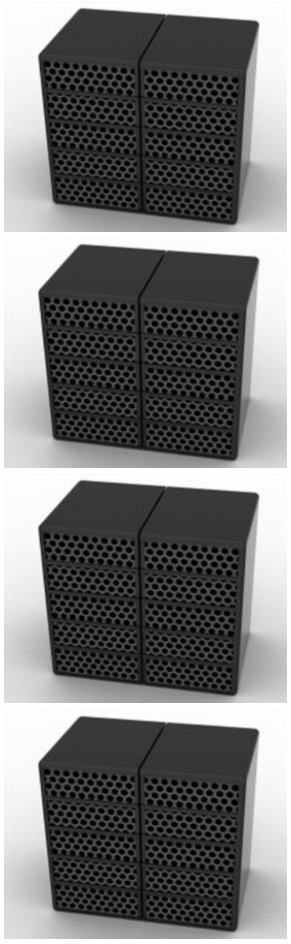
Most Variant Calling Pipelines Today

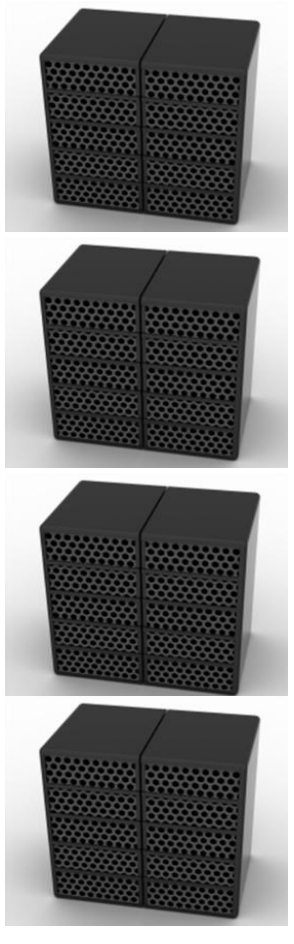


36-120 hrs from fastq to bam to vcf

Parallelism at the Sample Level

Data-Level Parallelism





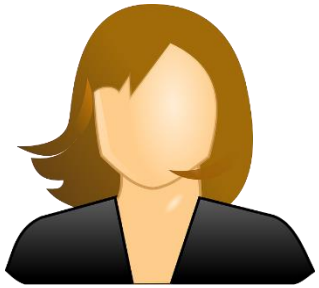
Read genome on 1 machine: ~1000 seconds*
Read genome on 1000 machines: ~1 second*

*112 GB, 35x coverage

What is Data Parallel Programming?

```
lines = sc.textFile(”jim.cram”)
lineLengths = lines.map(lambda s: len(s))
totalLength = lineLengths.reduce(lambda a, b: a + b)
```

What is Data Science?



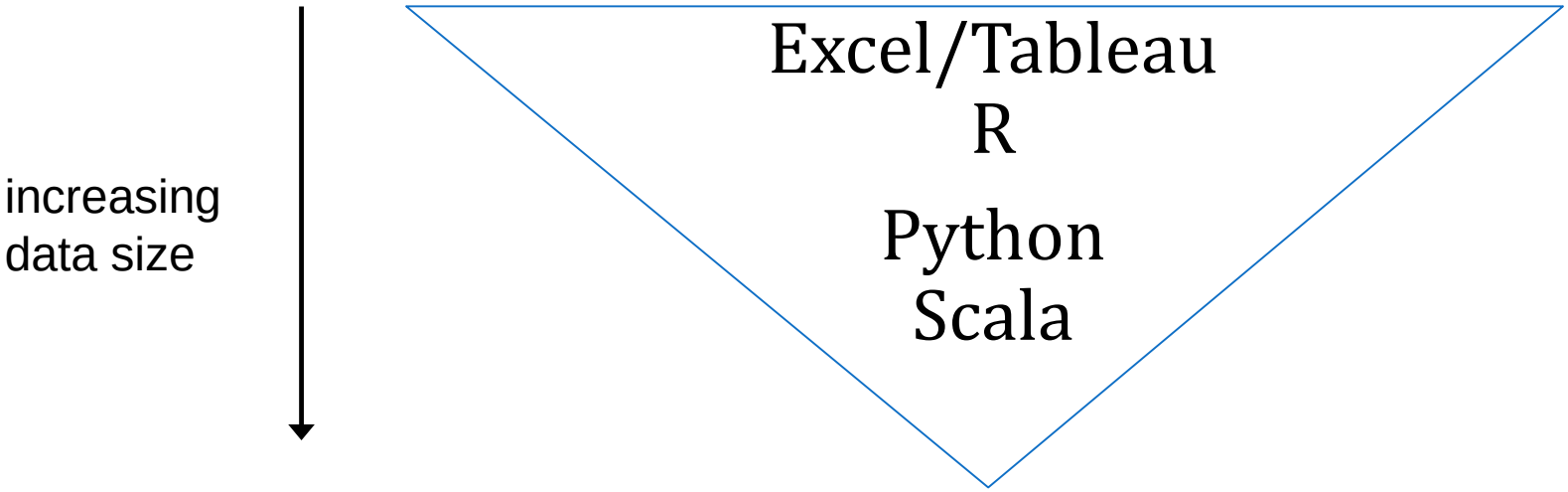
Software

Data

Gain insights of maximal value from data, while securely minimizing the cost needed to acquire said insights

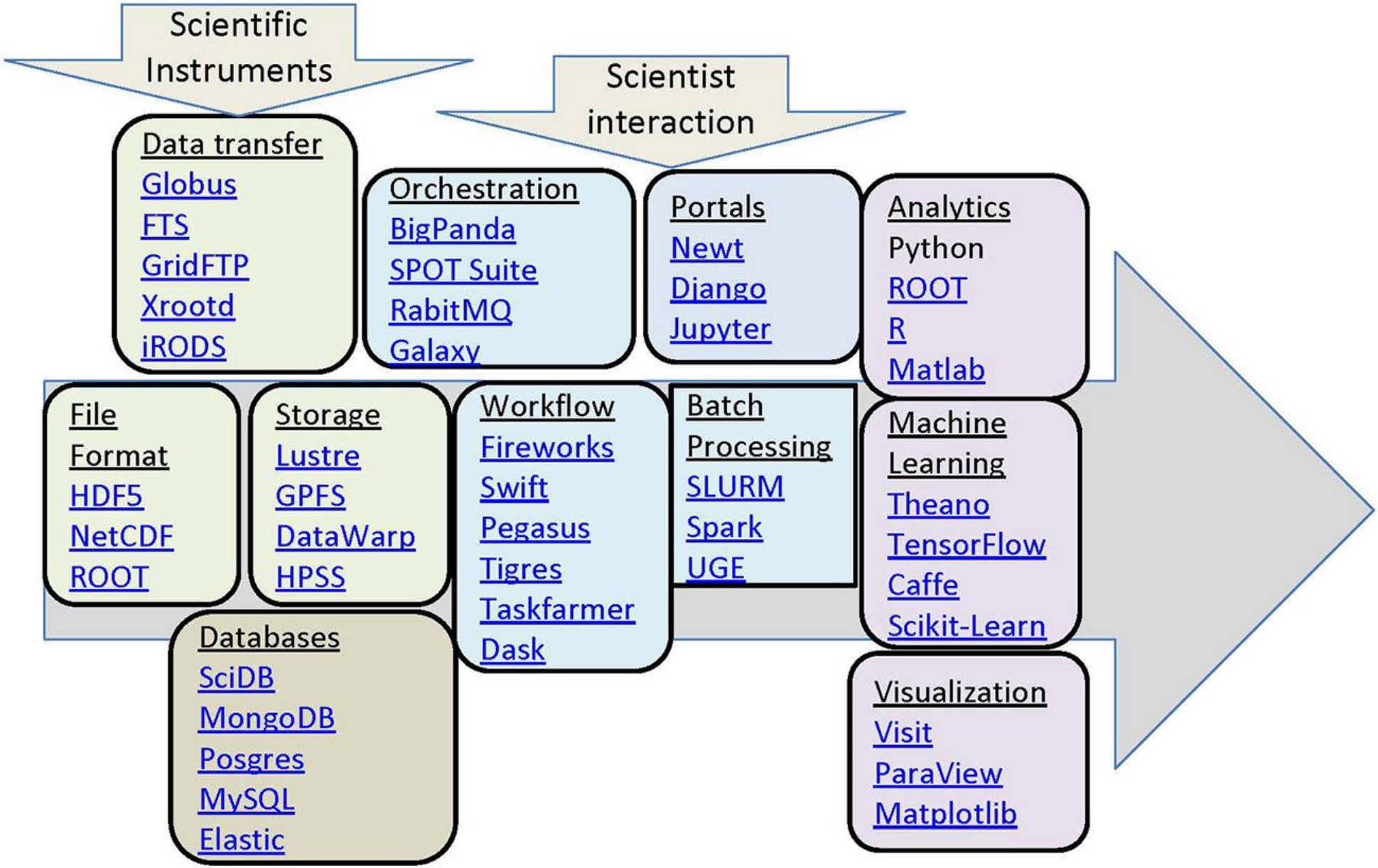
Insights

Programming for Data Science

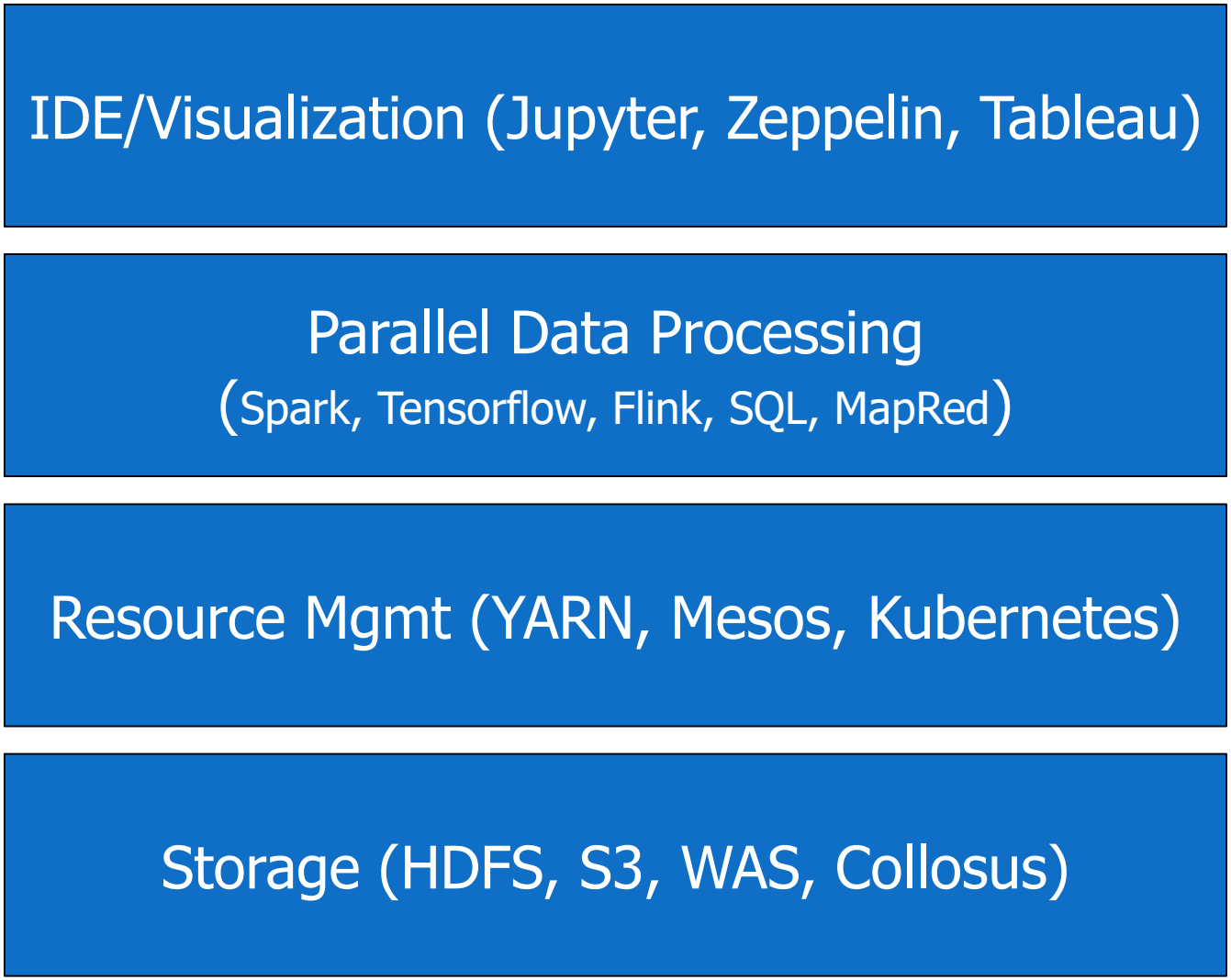


Tools: R/Studio, Jupyter, Zeppelin, Tableau

Tools for the E-Science Data Scientist

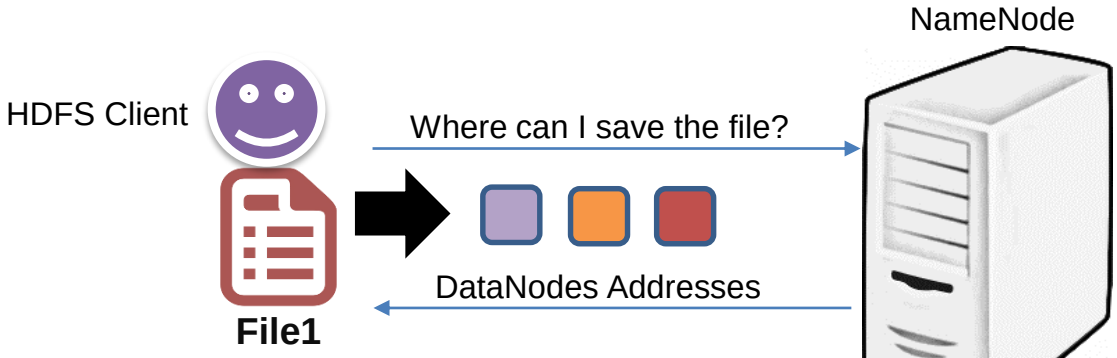


Tools for the Industrial Data Scientist



Hadoop

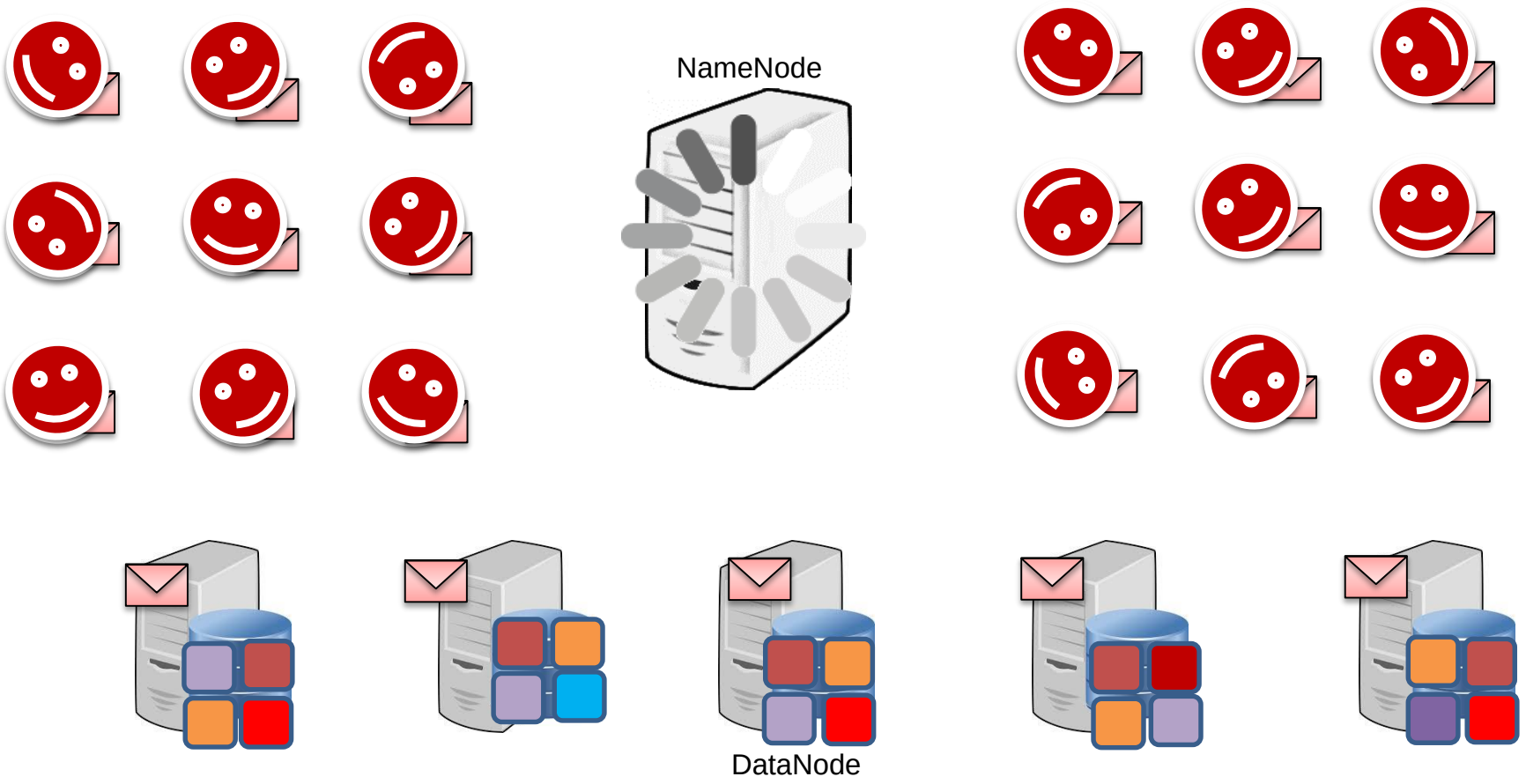
Hadoop Distributed Filesystem (HDFS)



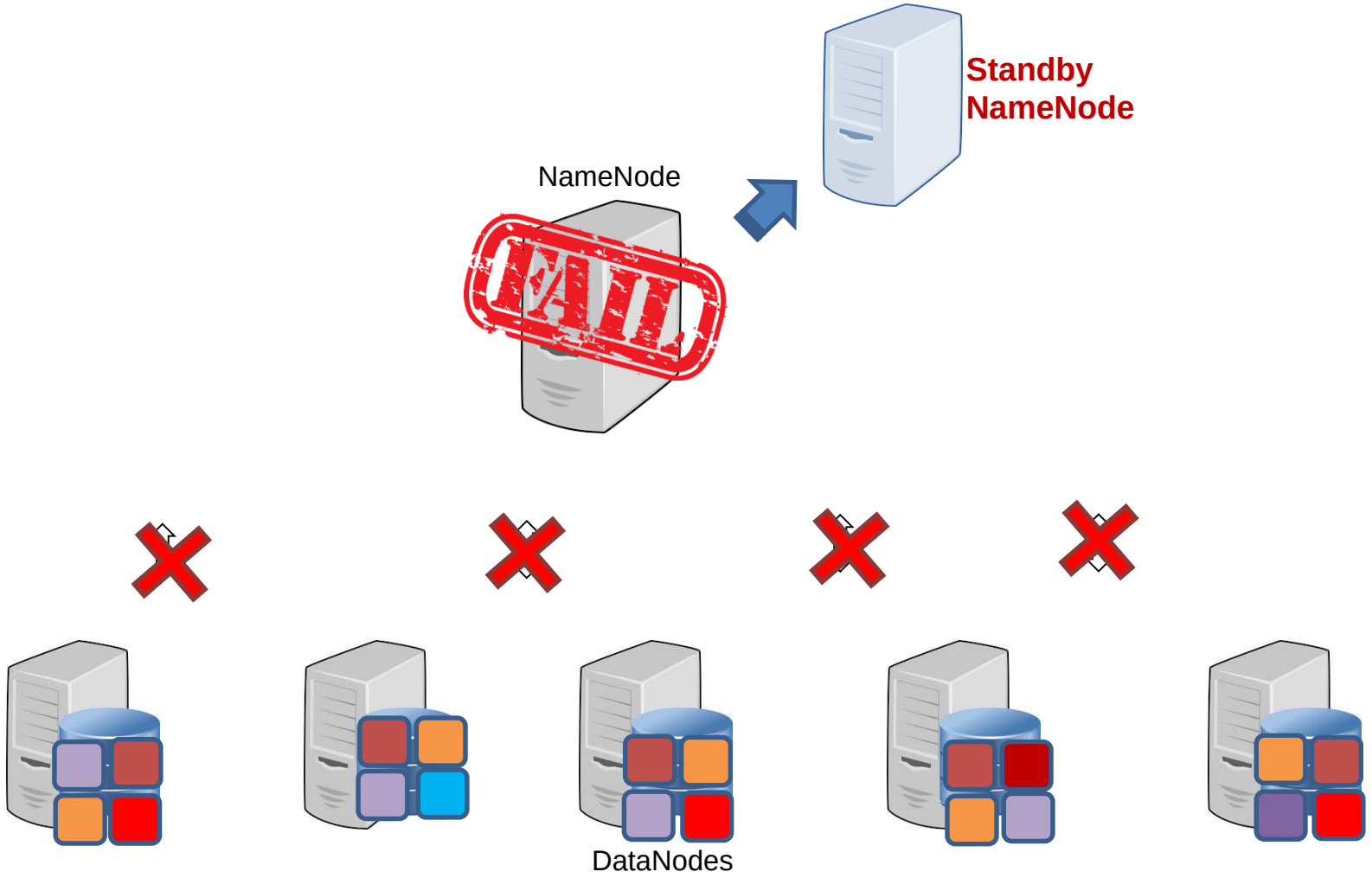
File System Metadata

File	Blocks Mappings
File1	Blk1 → DN1, Blk2 → DN5, Blk3 → DN3
File2	Blk1 → DN1, Blk2 → DN4
File3	Blk1 → DN1, Blk2 → DN2, Blk3 → DN3
File4	Blk1 → DN100
File5	Blk1 → DN4, Blk2 → DN2, Blk3 → DN9
...	... ...
FileN	Blk1 → DN2, Blk2 → DN8

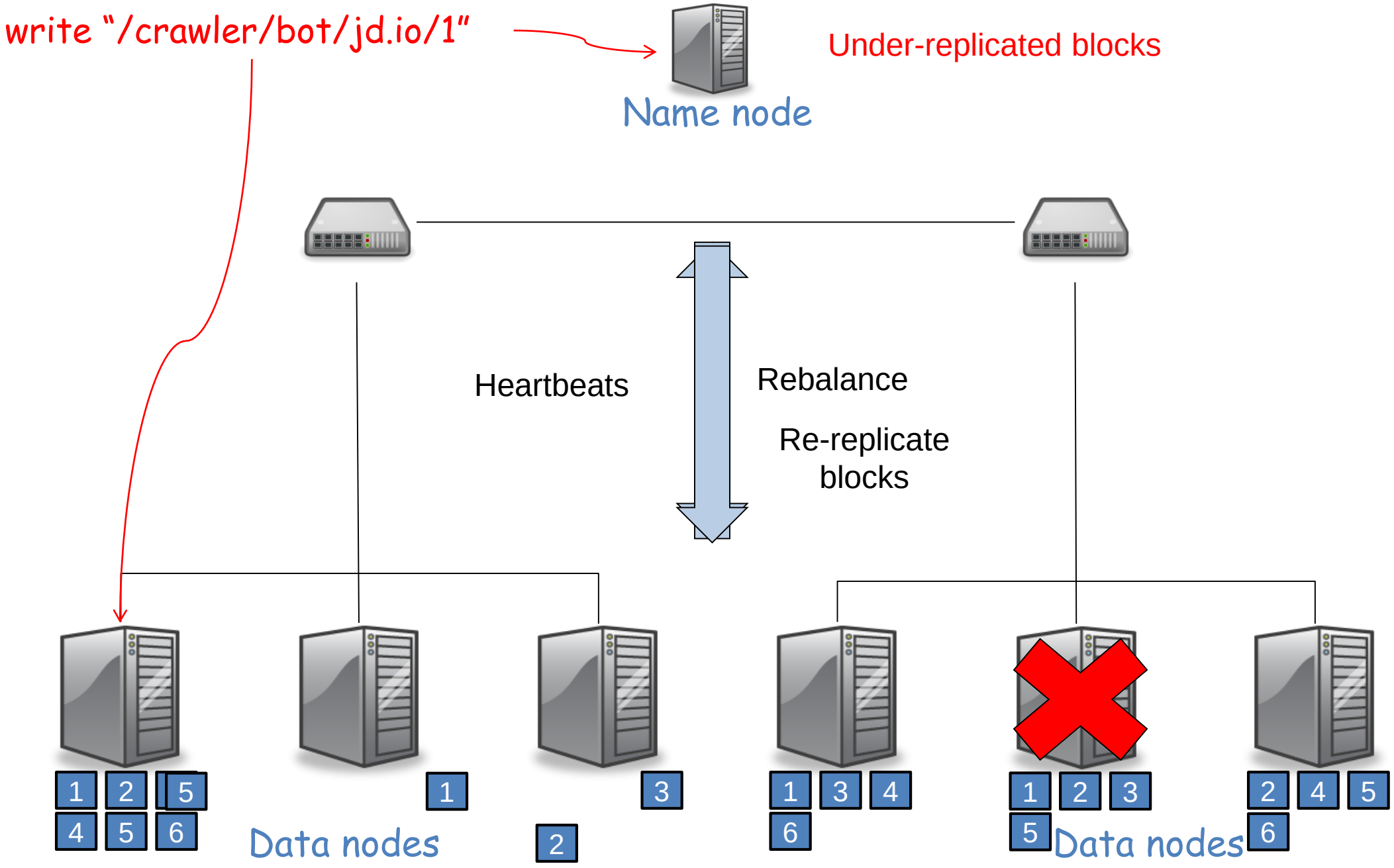
HDFS Performance At Scale



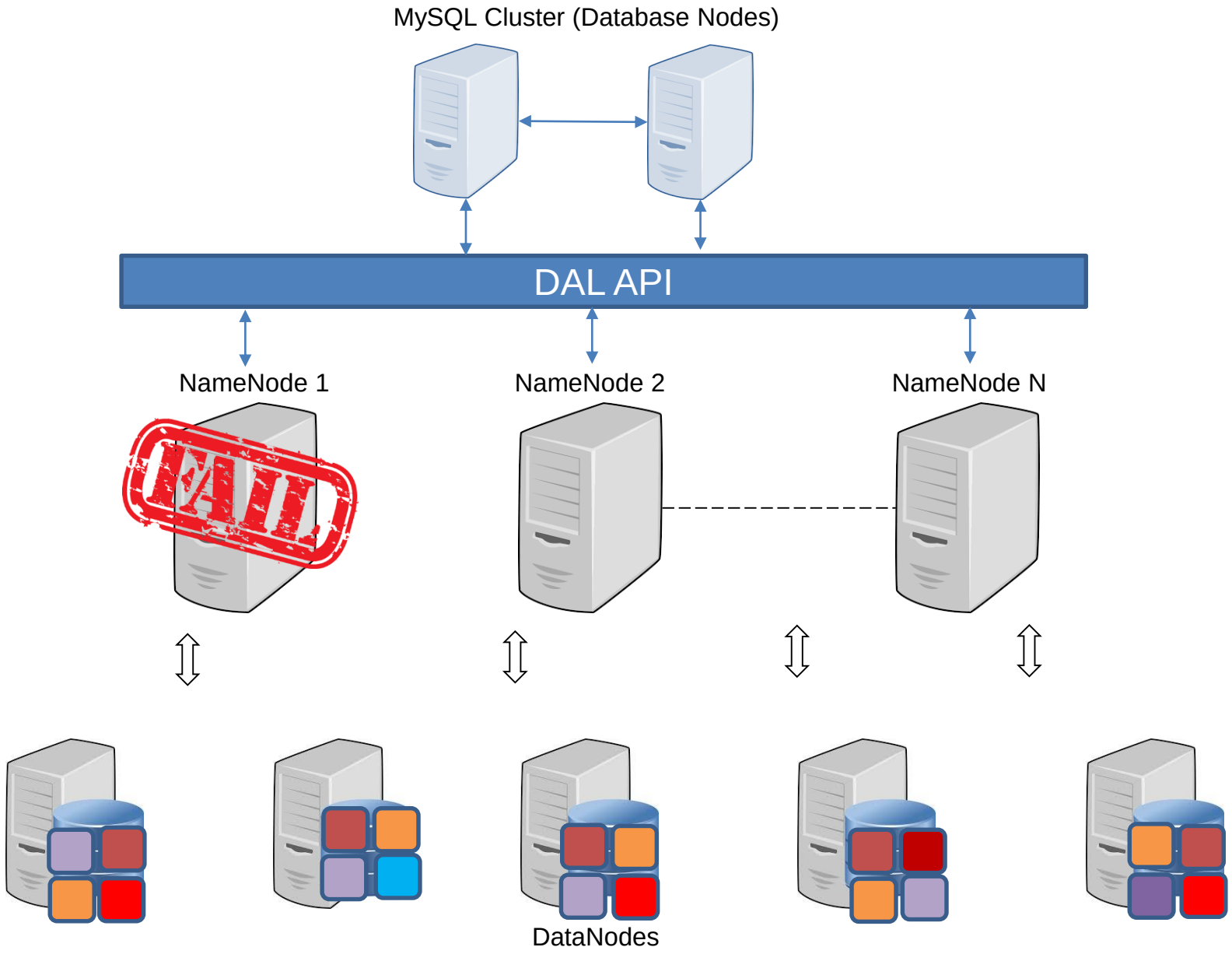
HDFS High-Availability



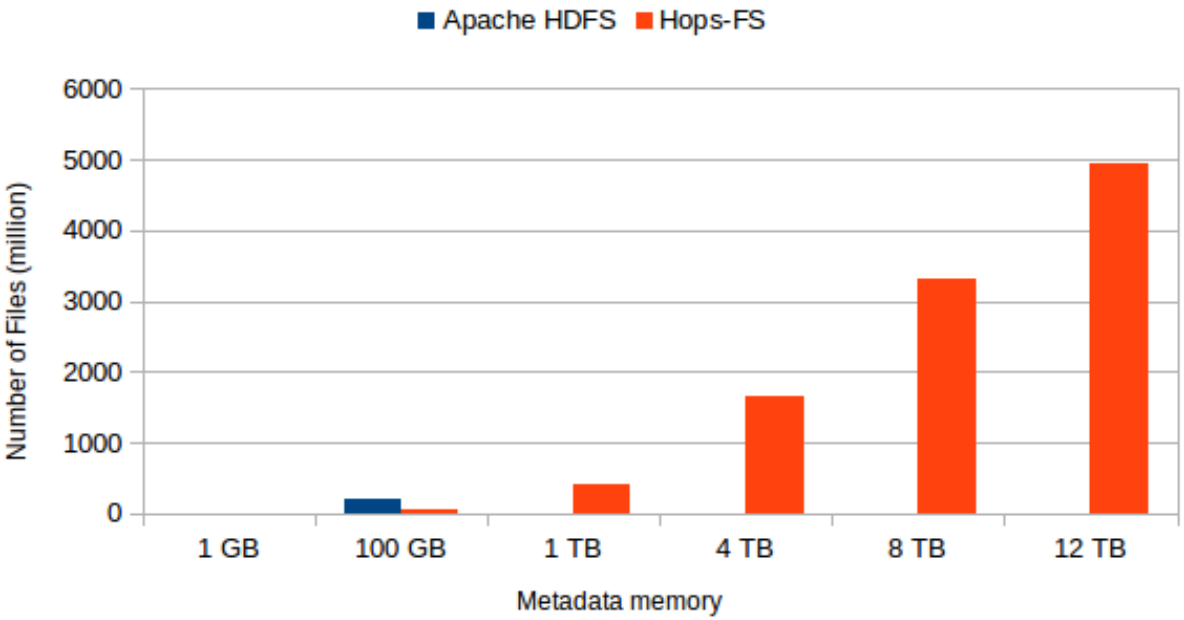
HDFS Protocols

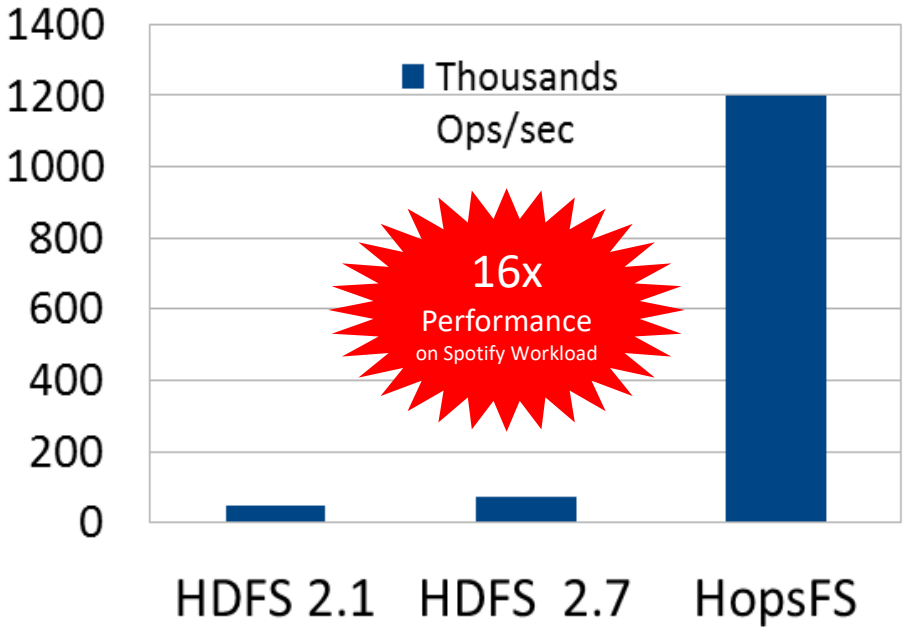


HopsFS


MySQL Cluster (Database Nodes)
DAL API
NameNode 1
NameNode 2
NameNode N
FAIL
DataNodes

HopsFS (www.hops.io)*





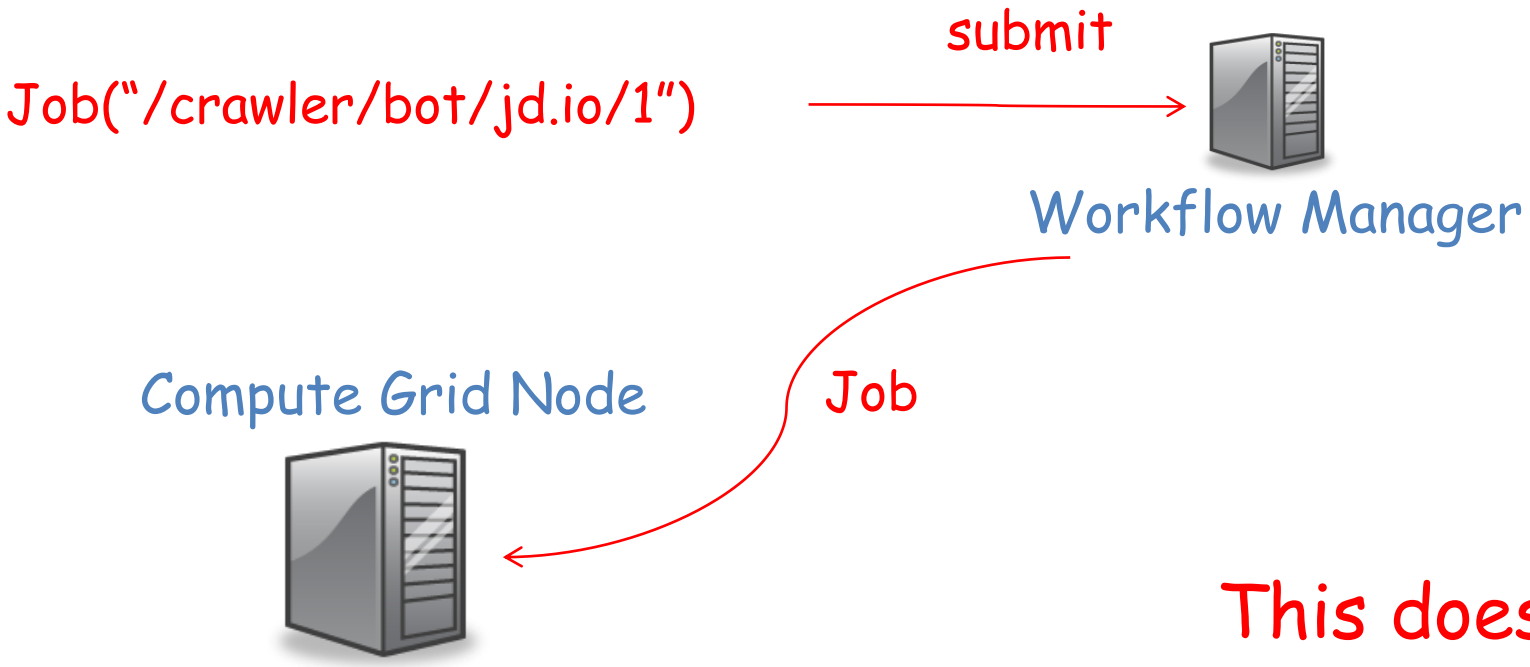
Bigger

Faster

*https://www.usenix.org/conference/fast17/technical-sessions/presentation/niazi

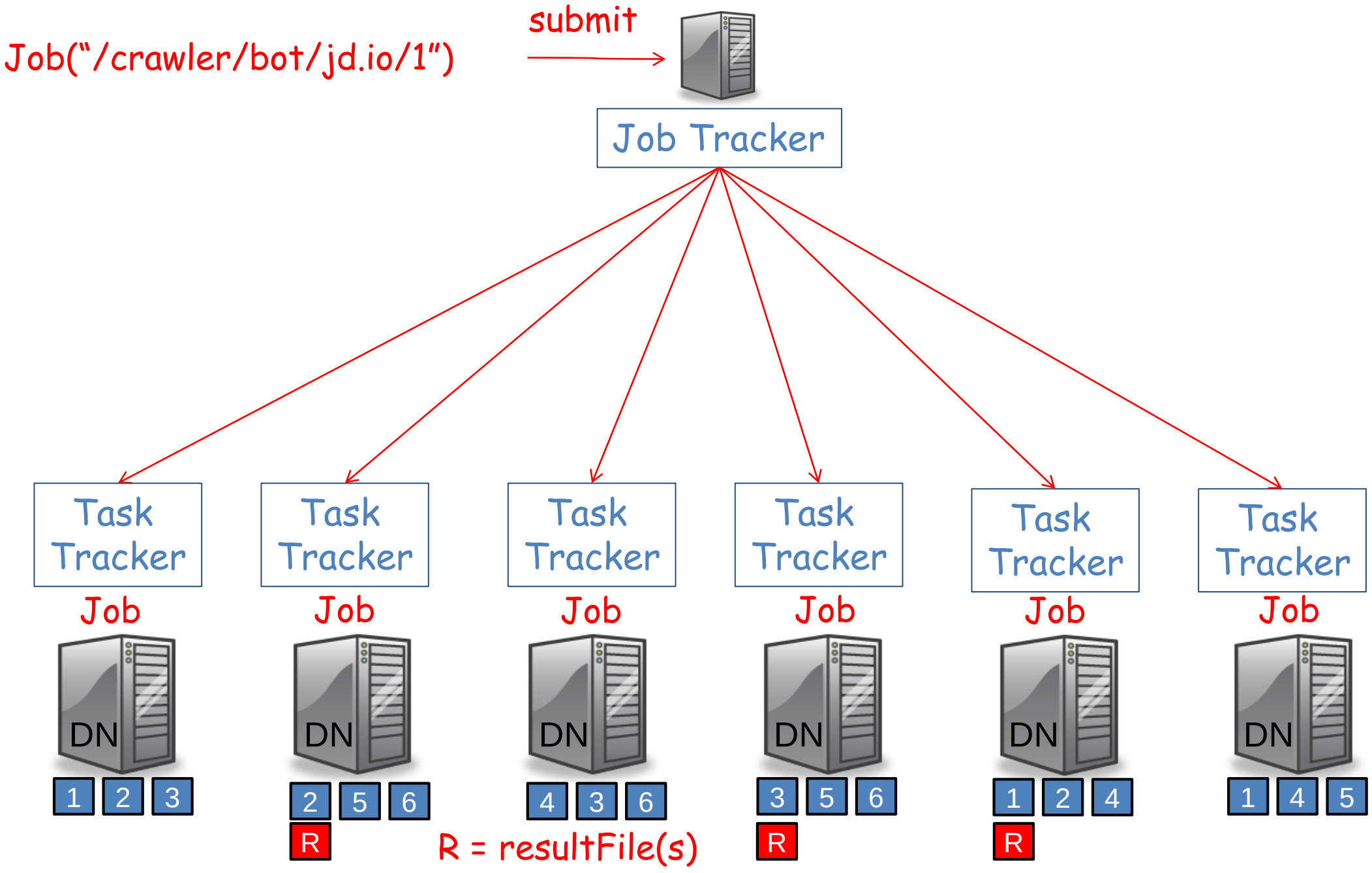
Processing Data in HDFS

Big Data Processing with No Data Locality

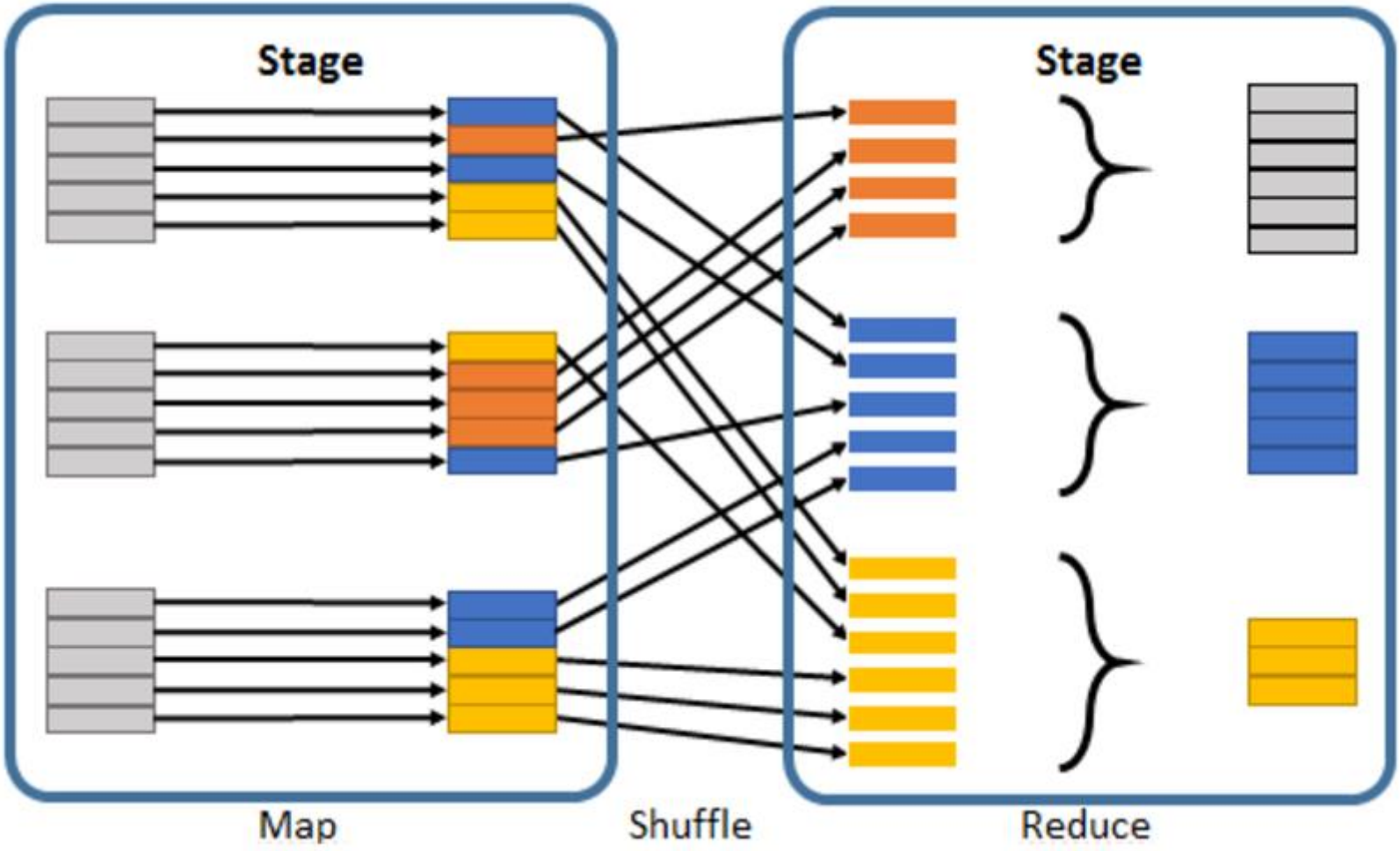


This doesn't scale.
Bandwidth is the bottleneck

MapReduce – Data Locality



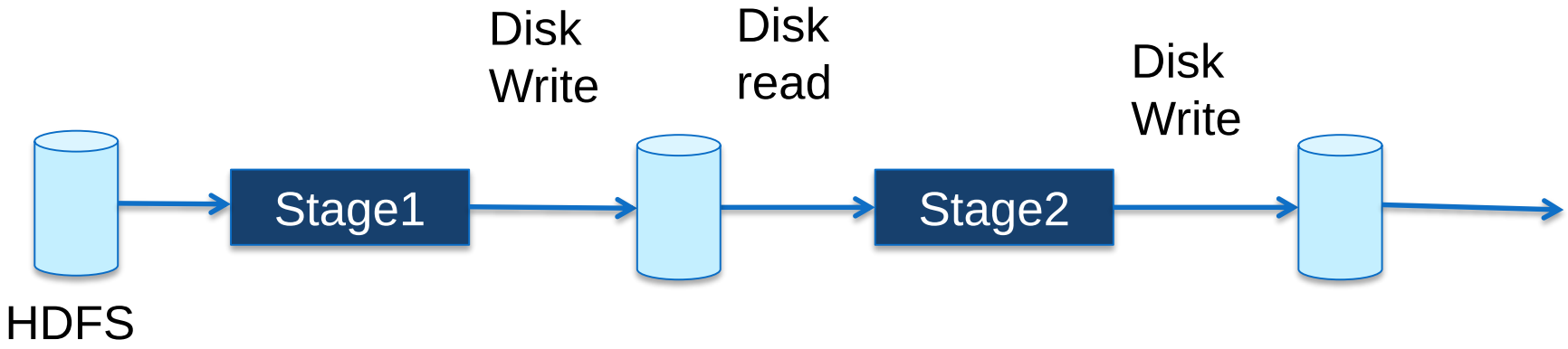
Stages in Spark



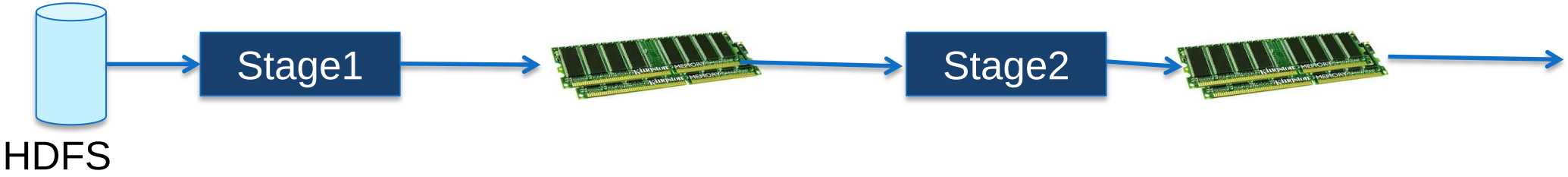
https://stackoverflow.com/questions/37528047/how-are-stages-split-into-tasks-in-spark

Spark Uses Memory instead of Disk

MapReduce: Persist results between stages to disk

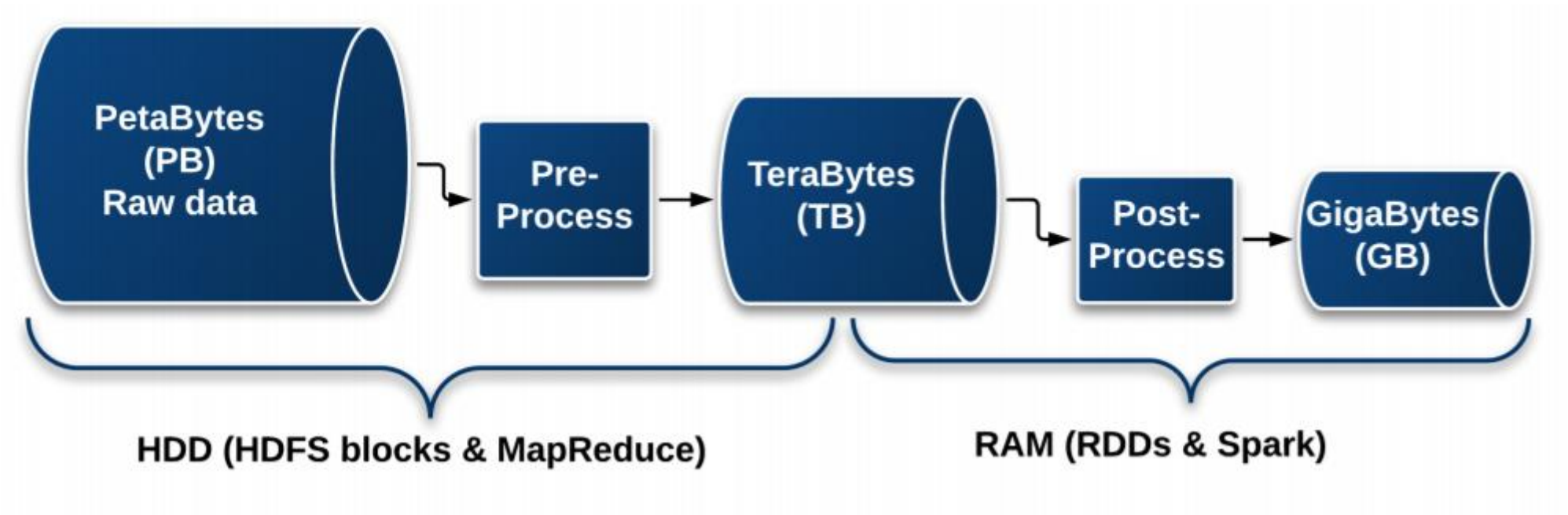


Spark: Store Results between Stages In-Memory



Genomics File Formats for Big Data

NextGen Processing Pipeline



Altti Ilari Maarala - Big Data Processing for Genomics, http://www.nordicehealth.se/wp-content/uploads/2016/12/BigDataGenomics.pdf

HDFS splits files into Blocks. Where is the Header?

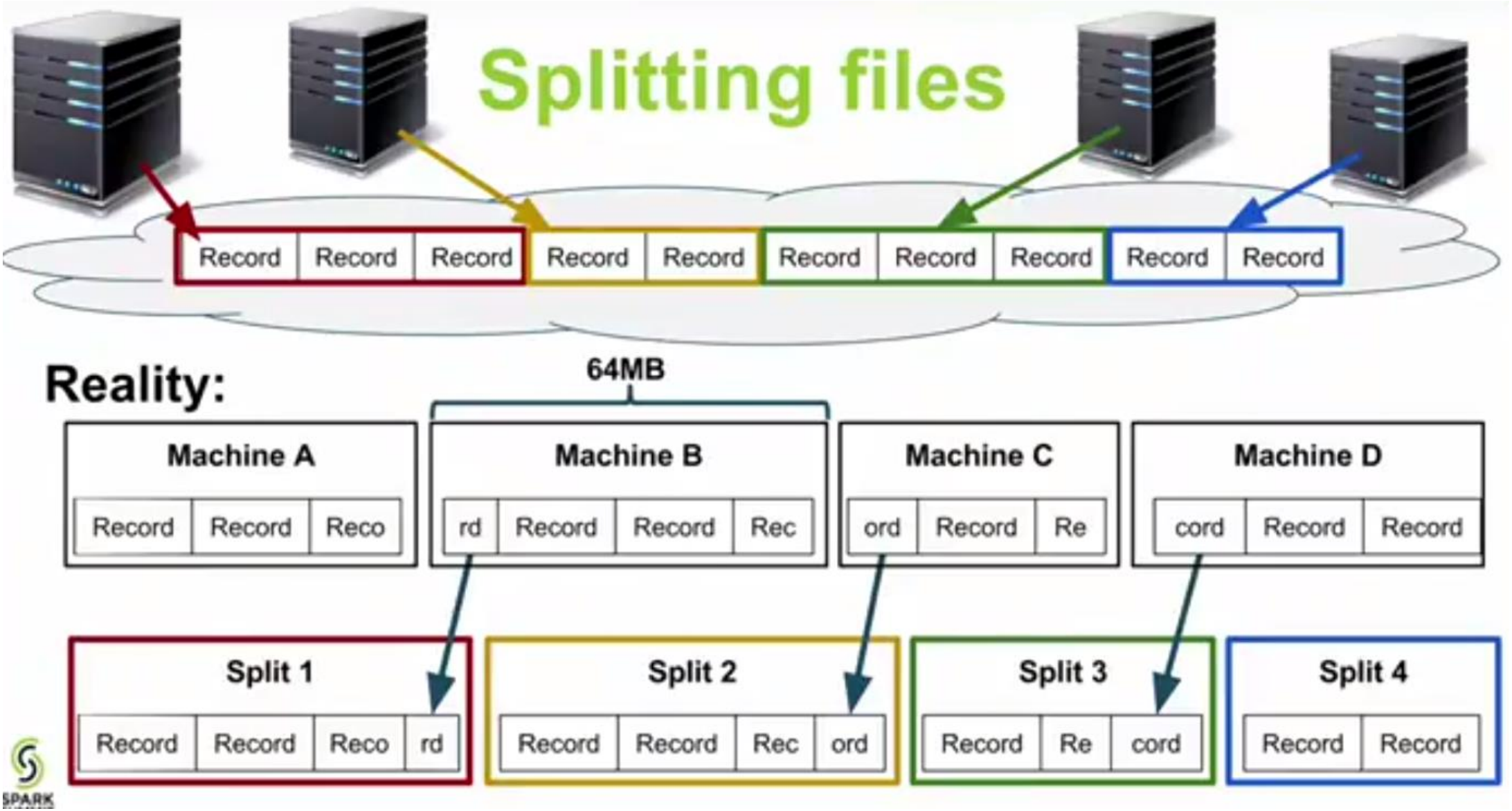


Image from [Ryan Williams, Spark Summit, 2017]

SAM Format

- Sequence Alignment/Map

Header

```
@HD  VN:1.4    GO:none   SO:coordinate
@SQ  SN:1 LN:249250621
@SQ  SN:2 LN:243199373
…
```

Reads

```
HWI-ST807:8592:79724   163   1   10001   0       101M     =   10009   109   TAACCCTAACC…
HWI-ST807:8592:79724    83   1   10009   0       101M     =   10001  -109   ACCCTAACCCT…
HWI-ST807:9505:89866   163   1   10048   29   20M1D81M     =   10368   374   CCAACCCTAAC…
HWI-ST807:6431:65669   163   1   10335   29    1S90M2D     =   10458   224   CAACCCTAACC…
…
```

We could split blocks in HDFS along newlines, but the data would not be compressed

Image from [Ryan Williams, Spark Summit, 2017]

BAM is block-based compression of SAM files

→ SAM format

\+ Binary record codec:

#bytes	contig	start	mapq	len(name)	name	len(cigar)	flags	len(seq)	cigar	seq	quals	tags

\+ Block-gzip compression (BGZF):

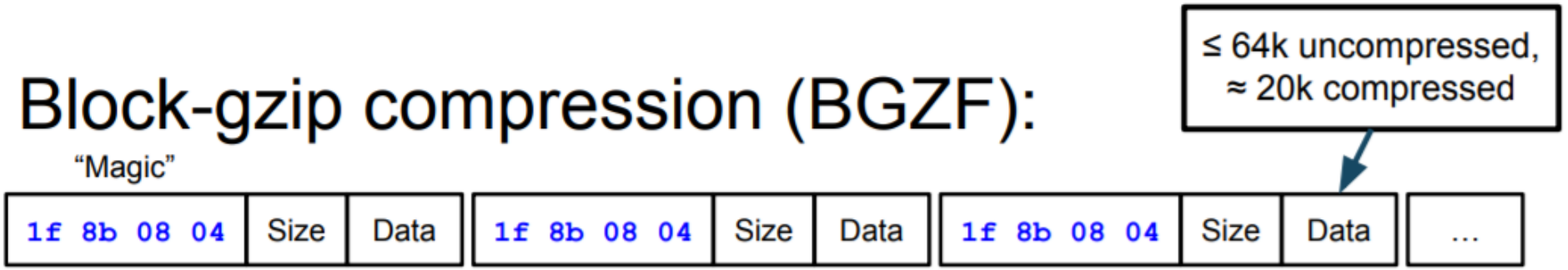


Image from [Ryan Williams, Spark Summit, 2017]

Hadoop BAM

- Helps split BAM files

  https://github.com/HadoopGenomics/Hadoop-BAM

- A library for processing NGS data formats in parallel with both Hadoop and Spark
  - Includes Hadoop I/O interface and tools for e.g., sorting, merging, filtering read alignments
  - Supported fileformats: BAM, SAM, CRAM, FASTQ, FASTA, QSEQ, BCF, and VCF
  - Used in GATK4, Adam, Halvade, SeqPig

”Hadoop-BAM: Directly Manipulating Next Generation Sequencing Data in the Cloud.” Niemenmaa, M., Kallio, A., Schumacher, A., Klemela, ¨ P., Korpelainen, E., and Heljanko, K. Bioinformatics 28(6):876-877, 2012.

Hadoop BAM – Low Level API

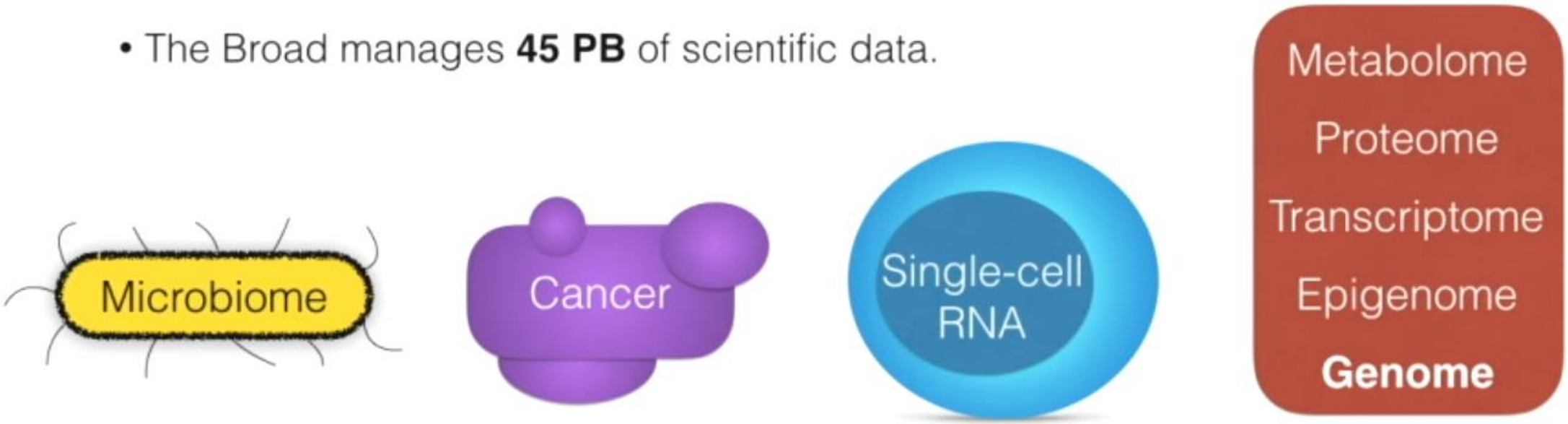
```
//Spark and SQL context initializations
SparkConf conf = new SparkConf();
JavaSparkContext sc = new JavaSparkContext(conf);
SQLContext sqlContext = new SQLContext(sc);//HiveContext if HiveQL used
//Reading BAM file into RDD from HDFS
1 JavaPairRDD<LongWritable, SAMRecordWritable> bamRDD =
sc.newAPIHadoopFile("alignments.bam", BAMInputFormat.class,
LongWritable.class, SAMRecordWritable.class, sc.hadoopConfiguration());
//Mapping to Serializable MyAlignment RDD
2 JavaRDD<MyAlignment> rdd = bamRDD.values().map(bam -> new MyAlignment
(bam.getReadName(), bam.getStart(), bam.getReadBases(),
bam.getReadUnmappedFlag()...));
//Create DataFrame and register table
3 DataFrame bamDF = sqlContext.createDataFrame(rdd, BAMRecord.class);
4 bamDF.registerTempTable("bamrecords");
//Filter unmapped reads and sort
5 DataFrame result = sqlContext.sql(
"SELECT * FROM bamrecords WHERE unmapped=true ORDER BY position ASC");

//Serializable class needed for DataFrame schema
public class MyAlignment implements Serializable {
    public MyAlignment(String readID, Integer position, String bases,
boolean unmapped ...}
```

Big Data Genomics Frameworks

Genomics is Big Data

Broad Institute data

- The Broad sequences **1 genome every 10 minutes**.
- The Broad generates **17 TB** of new genomes per day.
- The Broad manages **45 PB** of scientific data.



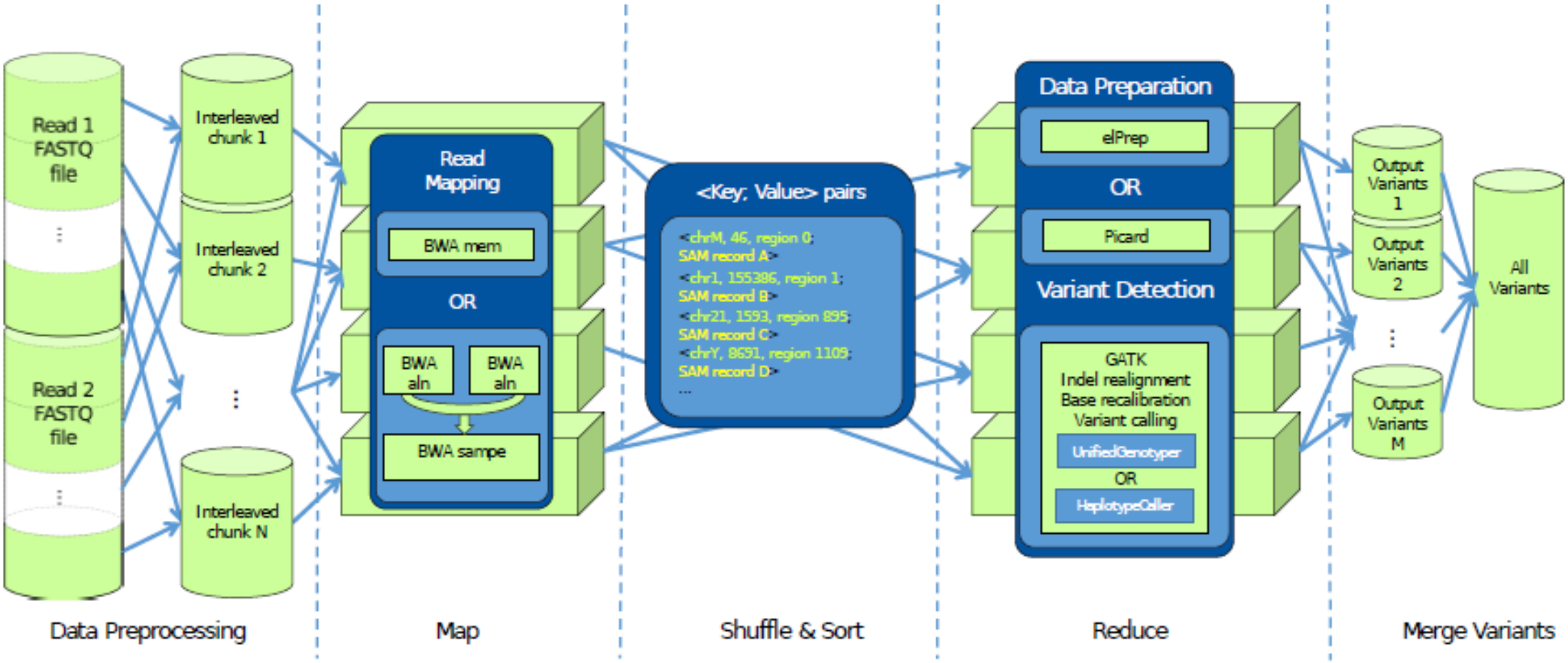
https://www.youtube.com/watch?v=pyeQusIN5Ao&list=PLlMMtlgw6qNjROoMNTBQjAcdx53kV50cS

Spark-based Genomic Analysis tools/platforms

- Broad Institute GATK4. **Spark/Scala**
  - Next generation of GATK suite of tools.
- Broad Institute Hail. **Spark/Python**
  - Variant analysis at scale
- RISE Lab (Berkeley) – ADAM. **Spark/Scala**
  - QC / variant-calling / viz tools
  - bdg-formats - avro schemas for genomic record-types
- Aalto HadoopBAM. **MapReduce/Java**
  - PigSeq, Metagenomics
- Halvade, Uni Ghent. **MapReduce/Java**
  - WGS Pipelines, RNASeq.
- SaaSFee, Humboldt Uni. **Cuneiform/YARN**
  - WGS Pipelines, RNASeq.

Halvade

- Runs existing GATK tools as MapReduce Jobs
  - Parallelizes reducers on the chromosome level (max 23 partitions)



[Figure 2.1, Phd Thesis, Dries Decap, Univ Ghent 2017]

Halvade Builds on Existing Tools

step	program	input	output
align reads	BWA	FASTQ	SAM
convert SAM to BAM	Picard	SAM	BAM
sort reads	Picard	BAM	BAM
mark duplicates	Picard	BAM	BAM
identify realignment intervals	GATK	BAM	intervals
realign intervals	GATK	BAM & intervals	BAM
build BQSR table	GATK	BAM	table
recalibrate base quality scores	GATK	BAM & table	BAM
call variants	GATK	BAM	VCF

[Table 2.1, Phd Thesis, Dries Decap, Univ Ghent 2017]

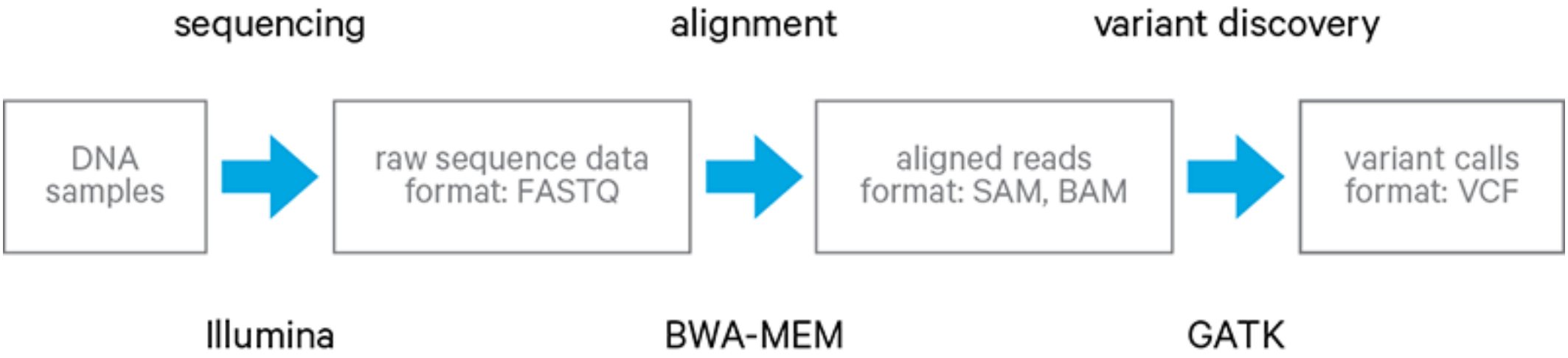
Halvade Benchmarks

Cluster	# worker nodes	# parallel tasks	# CPU cores	runtime
Intel Big Data cluster	1	3	18	47h 59min
	4	15	90	9h 54min
	8	31	186	4h 50min
	15	59	354	2h 39min
Amazon EMR	1	4	32	38h 38min
	2	8	64	20h 19min
	4	16	128	10h 20min
	8	32	256	5h 13min
	16	64	512	2h 44min

[Table 2.2, Phd Thesis, Dries Decap, Univ Ghent 2017]

HAIL

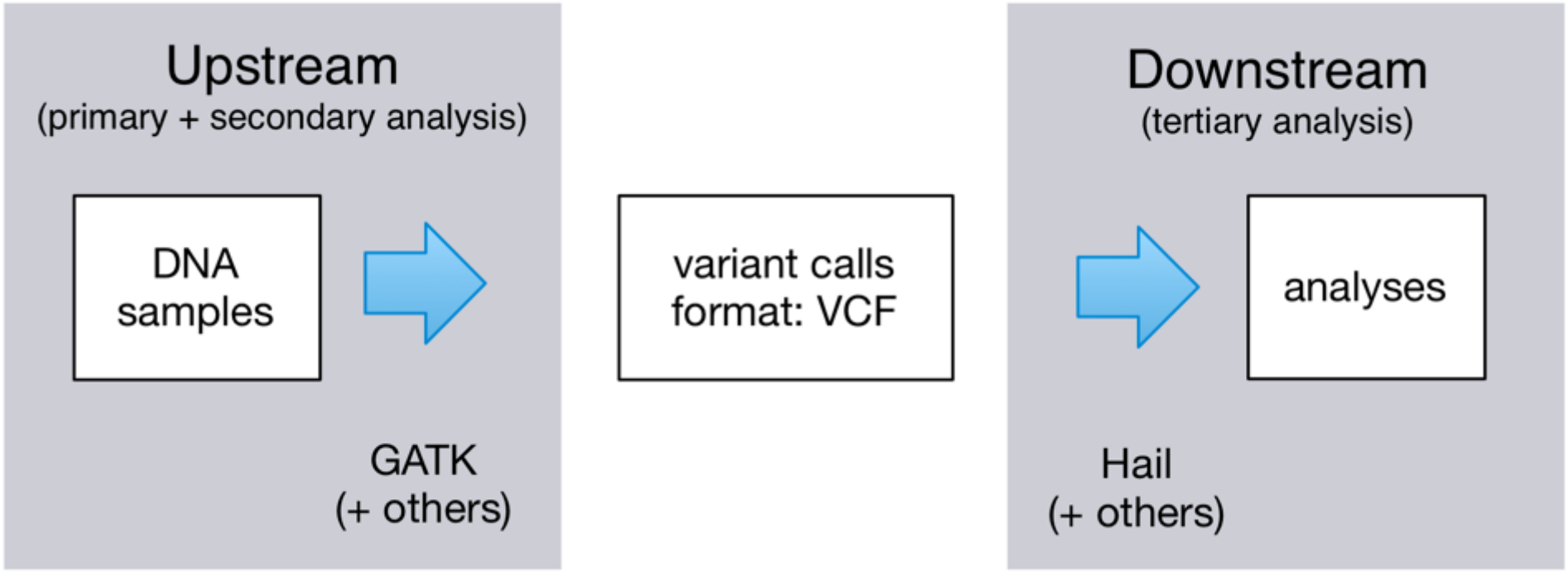
- Hail was written from the outset to use Apache Spark so it could take advantage of the ability to scale to thousands of nodes and petabytes of data. Hail is released under the MIT open source license



http://blog.cloudera.com/blog/2016/04/genome-analysis-toolkit-now-using-apache-spark-for-data-processing/

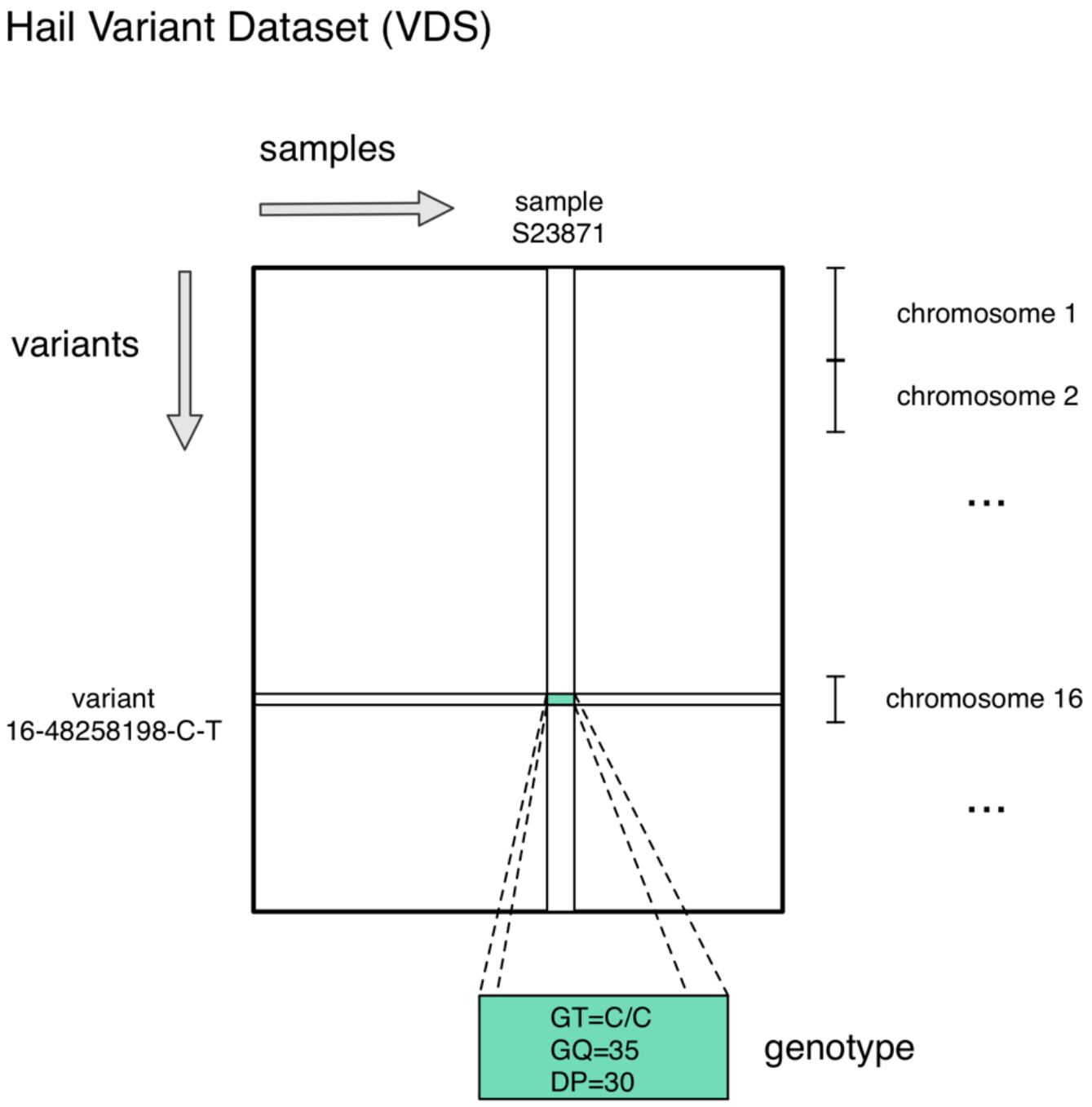
HAIL

- Hail was used for running QC and basic analysis of over 120,000 exomes and 15,000 genomes for creating the Genome Aggregation Database (gnomAD)

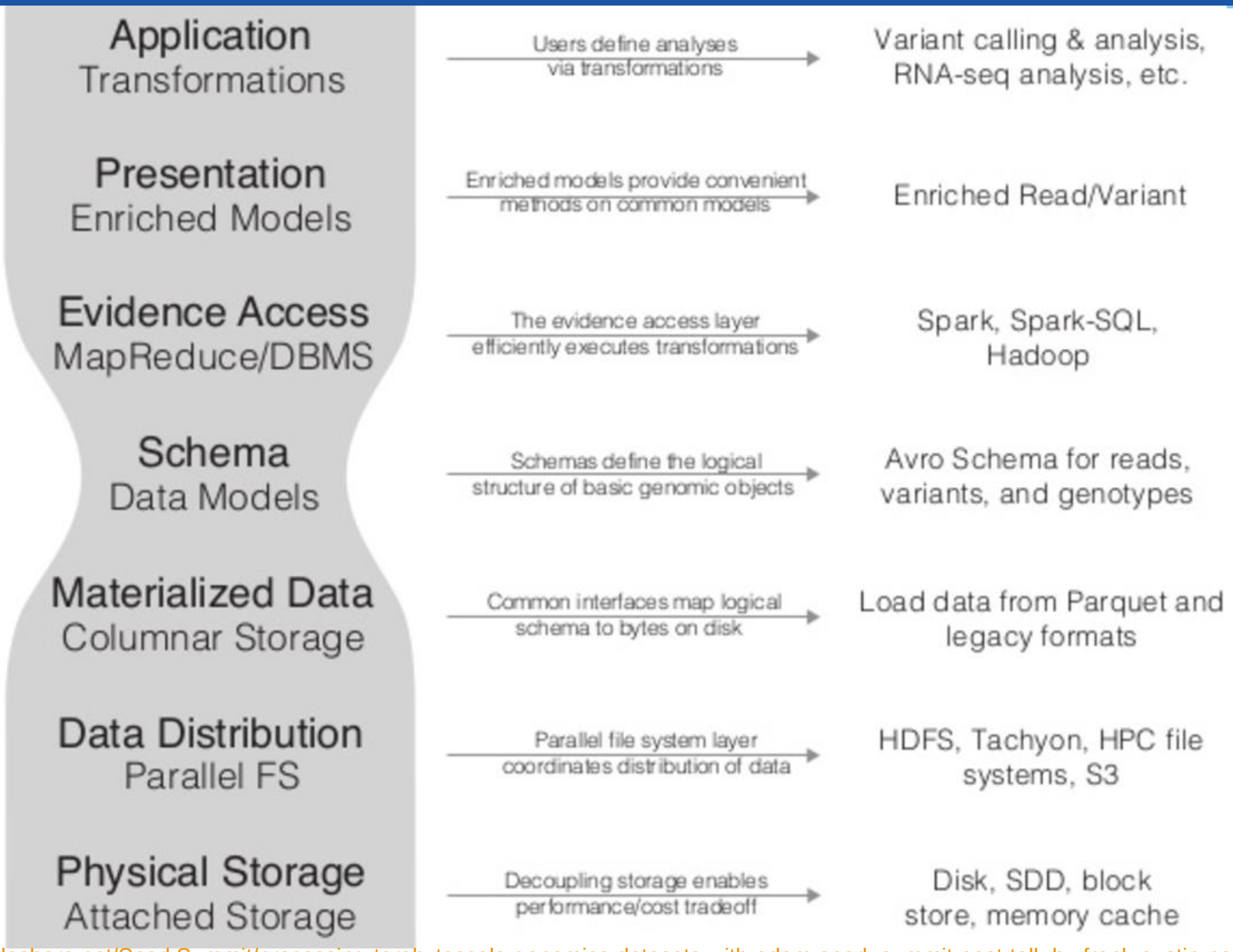


https://blog.cloudera.com/blog/2017/05/hail-scalable-genomics-analysis-with-spark/

Hail Variant Dataset



ADAM



https://www.slideshare.net/SparkSummit/processing-terabytescale-genomics-datasets-with-adam-spark-summit-east-talk-by-frank-austin-nothaft

ADAM

```
record AlignmentRecord {
    union { null, Contig } contig = null;
    union { null, long } start = null;
    union { null, long } end = null;
    union { null, int } mapq = null;
    union { null, string } readName = null;
    union { null, string } sequence = null;
    union { null, string } mateReference = null;
    union { null, long } mateAlignmentStart = null;
    union { null, string } cigar = null;
    union { null, string } qual = null;
    union { null, string } recordGroupName = null;
    union { int, null } basesTrimmedFromStart = 0;
    union { int, null } basesTrimmedFromEnd = 0;
    union { boolean, null } readPaired = false;
    union { boolean, null } properPair = false;
    union { boolean, null } readMapped = false;
    union { boolean, null } mateMapped = false;
    union { boolean, null } firstOfPair = false;
    union { boolean, null } secondOfPair = false;
    union { boolean, null } failedVendorQualityChecks = false;
    union { boolean, null } duplicateRead = false;
    union { boolean, null } readNegativeStrand = false;
    union { boolean, null } mateNegativeStrand = false;
    union { boolean, null } primaryAlignment = false;
    union { boolean, null } secondaryAlignment = false;
    union { boolean, null } supplementaryAlignment = false;
    union { null, string } mismatchingPositions = null;
    union { null, string } origQual = null;
    union { null, string } attributes = null;
    union { null, string } recordGroupSequencingCenter = null;
    union { null, string } recordGroupDescription = null;
    union { null, long } recordGroupRunDateEpoch = null;
    union { null, string } recordGroupFlowOrder = null;
    union { null, string } recordGroupKeySequence = null;
    union { null, string } recordGroupLibrary = null;
    union { null, int } recordGroupPredictedMedianInsertSize = null;
    union { null, string } recordGroupPlatform = null;
    union { null, string } recordGroupPlatformUnit = null;
    union { null, string } recordGroupSample = null;
    union { null, Contig } mateContig = null;
}
```

Application
Transformations

Presentation
Enriched Models

Evidence Access
MapReduce/DBMS

Schema
Data Models

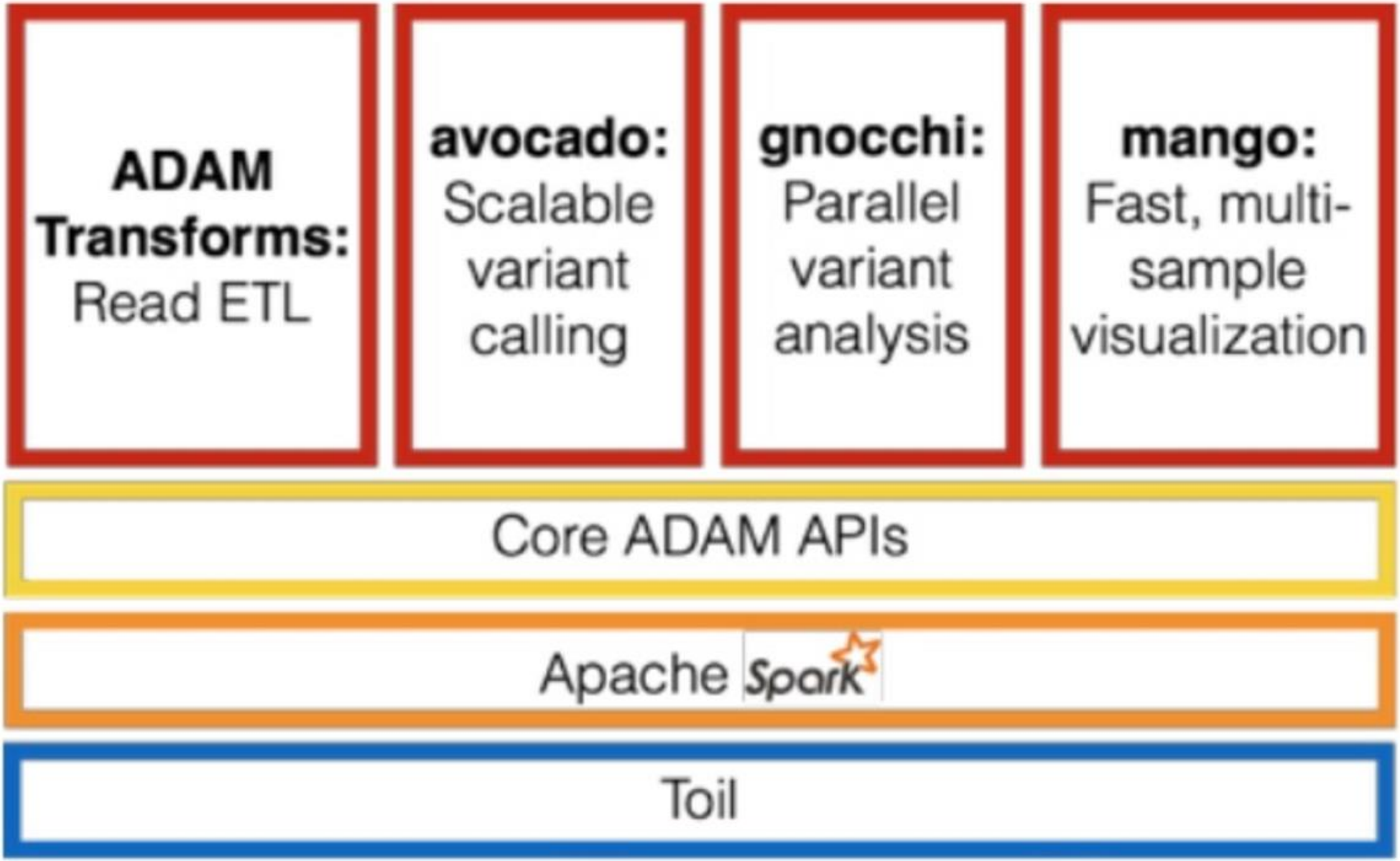
Materialized Data
Columnar Storage

Data Distribution
Parallel FS

Physical Storage
Attached Storage

https://www.slideshare.net/SparkSummit/processing-terabytescale-genomics-datasets-with-adam-spark-summit-east-talk-by-frank-austin-nothaft

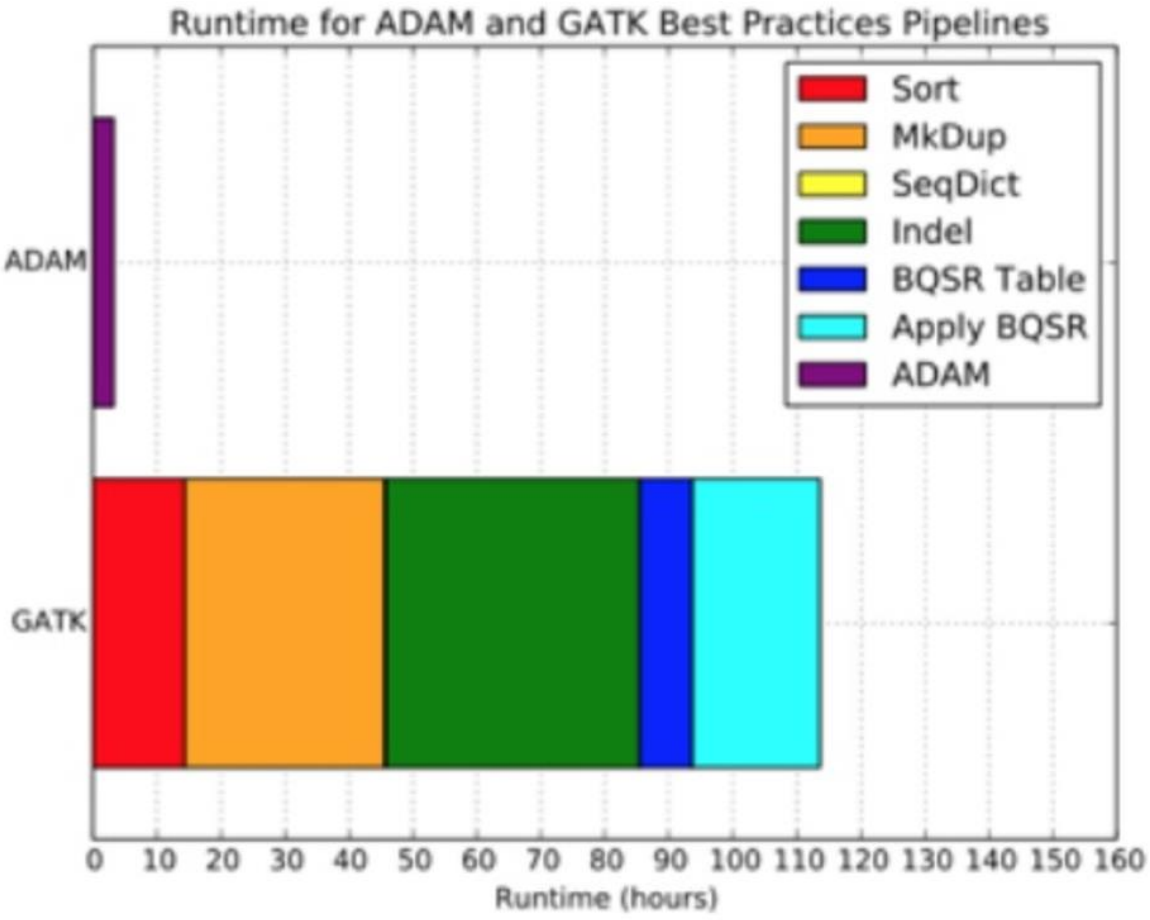
ADAM



https://www.slideshare.net/SparkSummit/processing-terabytescale-genomics-datasets-with-adam-spark-summit-east-talk-by-frank-austin-nothaft

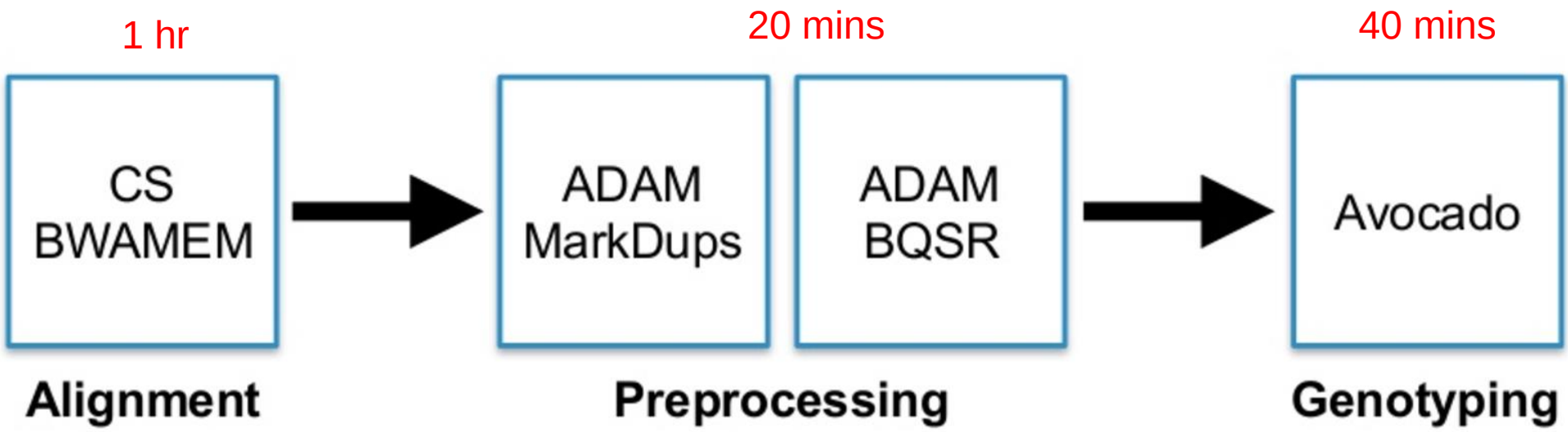
ADAM Performance

- ADAM produces statistically equivalent results to the GATK best practices pipeline
- Read preprocessing is >30x faster and 3x cheaper



https://www.slideshare.net/SparkSummit/processing-terabytescale-genomics-datasets-with-adam-spark-summit-east-talk-by-frank-austin-nothaft

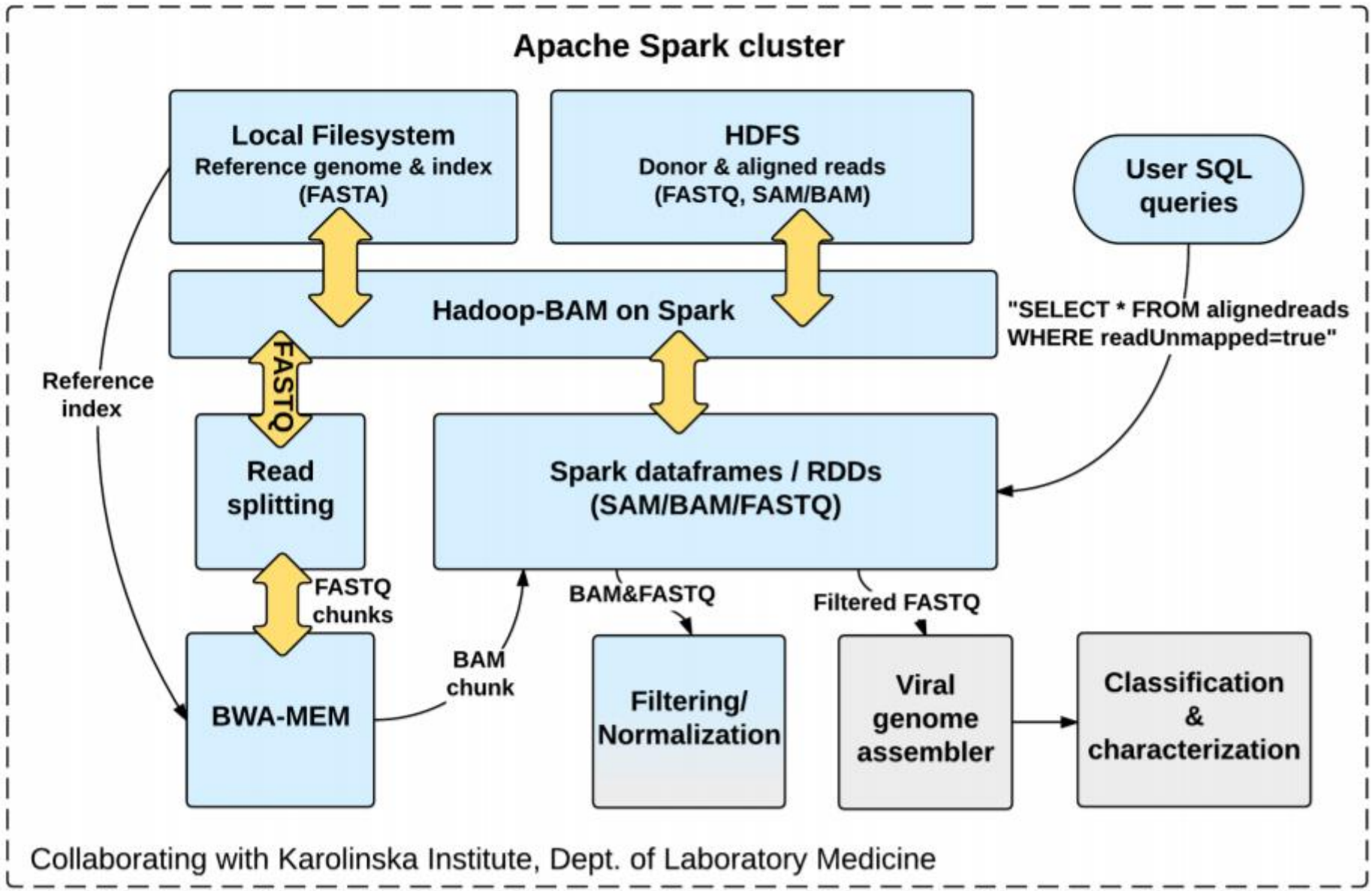
End-to-End Variant Analysis in Spark



- Can process a 65x whole genome in <2hrs on 1,024 cores
- CS-BWAMEM: https://github.com/ytchen0323/cloud-scale-bwamem

https://www.slideshare.net/SparkSummit/processing-terabytescale-genomics-datasets-with-adam-spark-summit-east-talk-by-frank-austin-nothaft

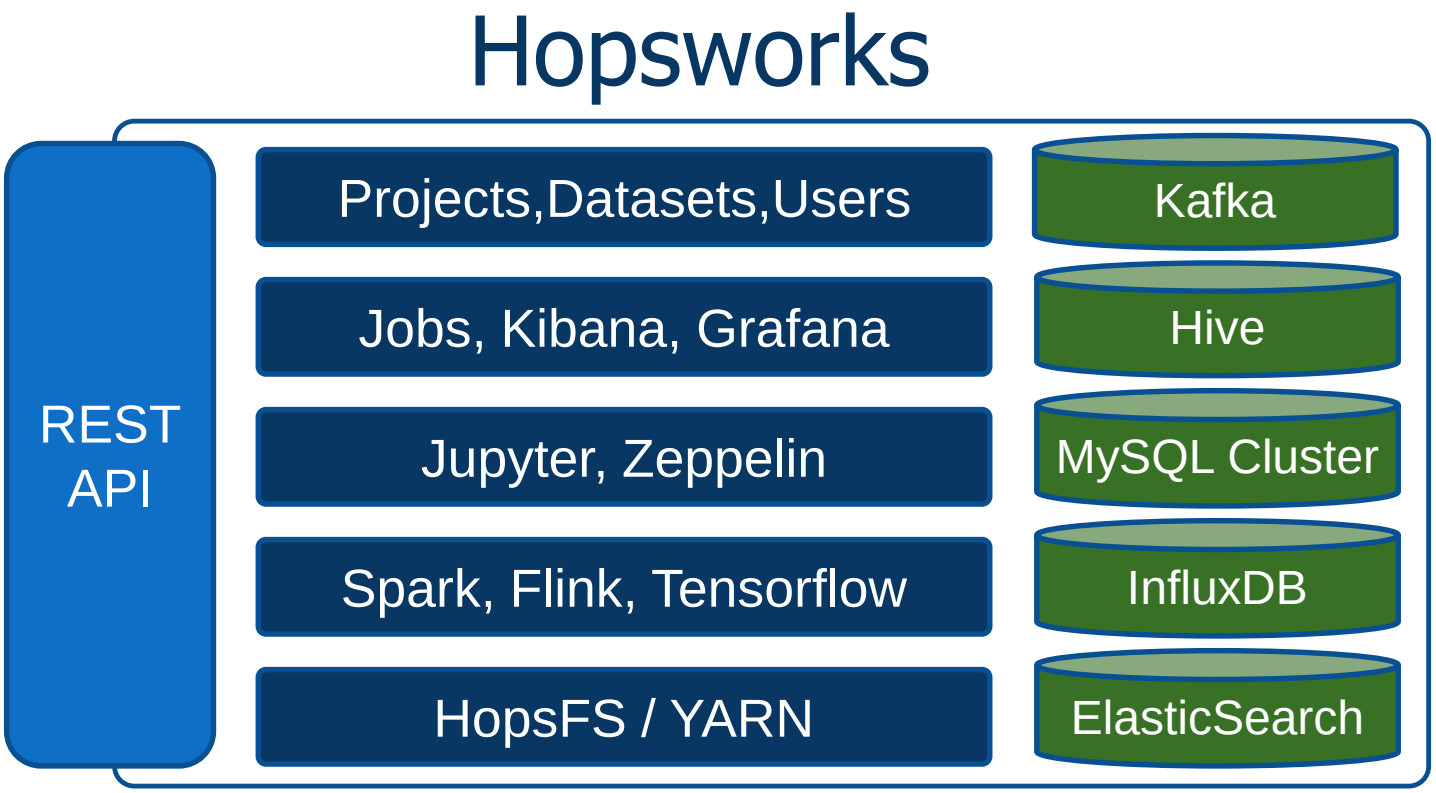
Parallel Pipelines for Metagenomics



Altti Ilari Maarala - Big Data Processing for Genomics, http://www.nordicehealth.se/wp-content/uploads/2016/12/BigDataGenomics.pdf

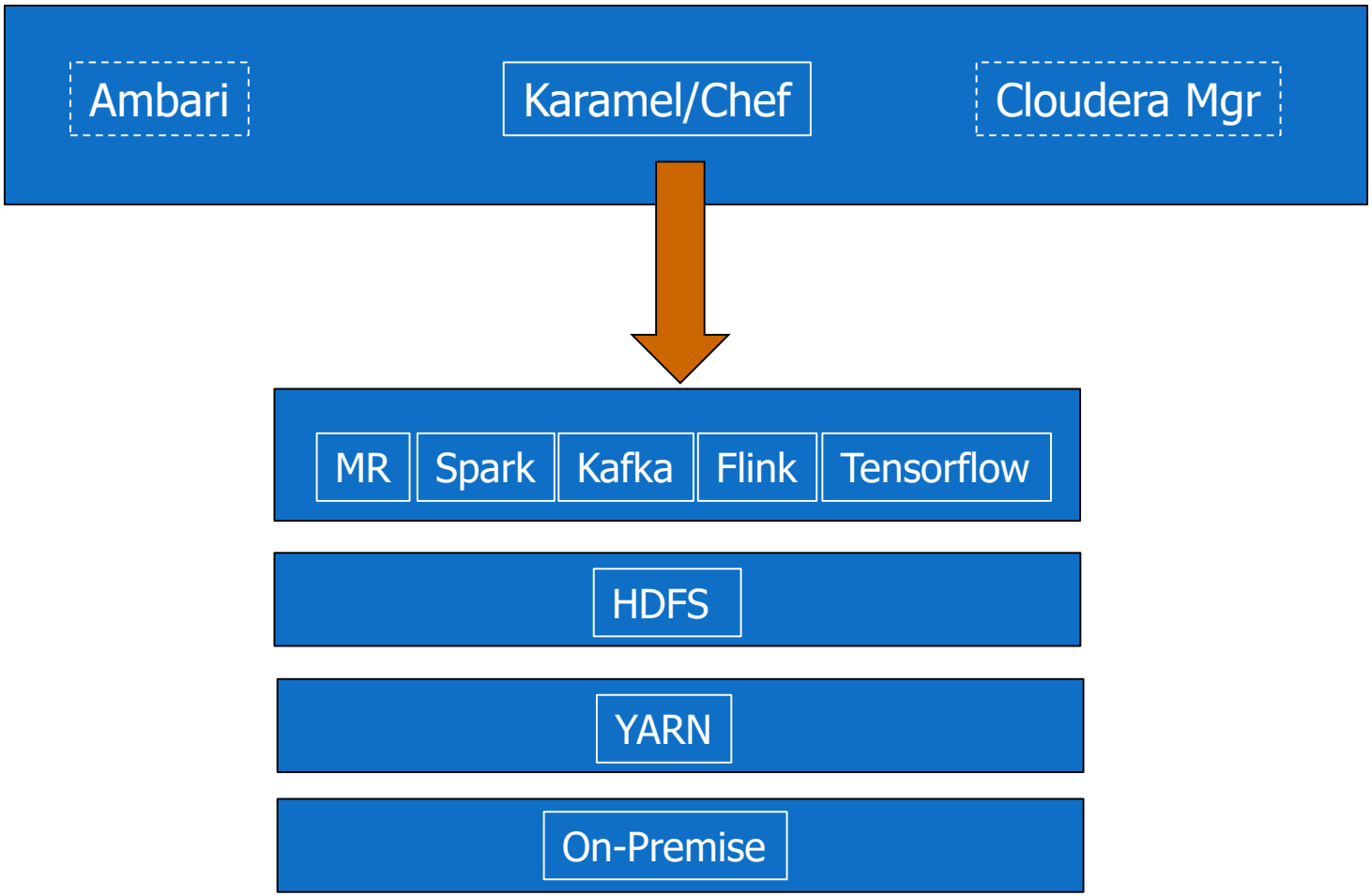
Hops Hadoop

Hopsworks Data Platform



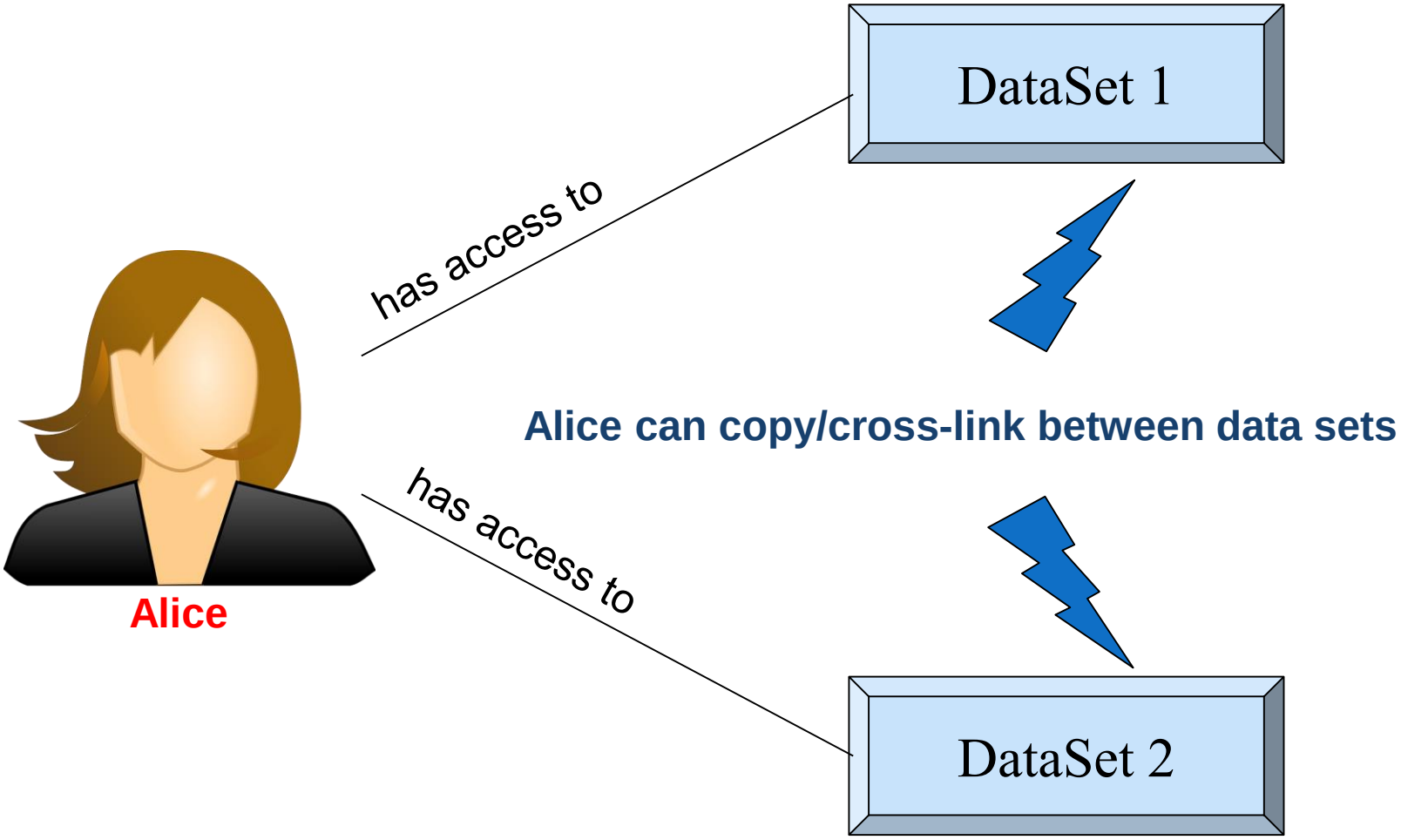
Hadoop Distributions Simplify Deployment

Install / Upgrade



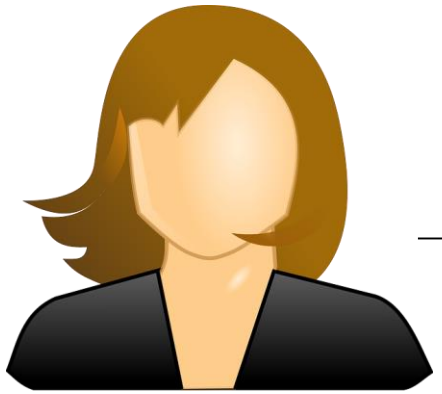
Hops is the only Hadoop Platform with end-to-end support for sensitive data

Cannot isolate Dataset Access in Hadoop



In Hadoop, only one Kerberos Identity is supported – no Dynamic Roles.

Hops Solution: Project-Specific UserIDs



NSA__Alice

Member of

Project 1

HDFS enforces access control

Users__Alice

Member of

Project 2

Sharing DataSets with Hops

NSA__Alice

Member of

Project 1

Add members of Project 1
to the DataSet X group

Users__Alice

Member of

Project 2

owns

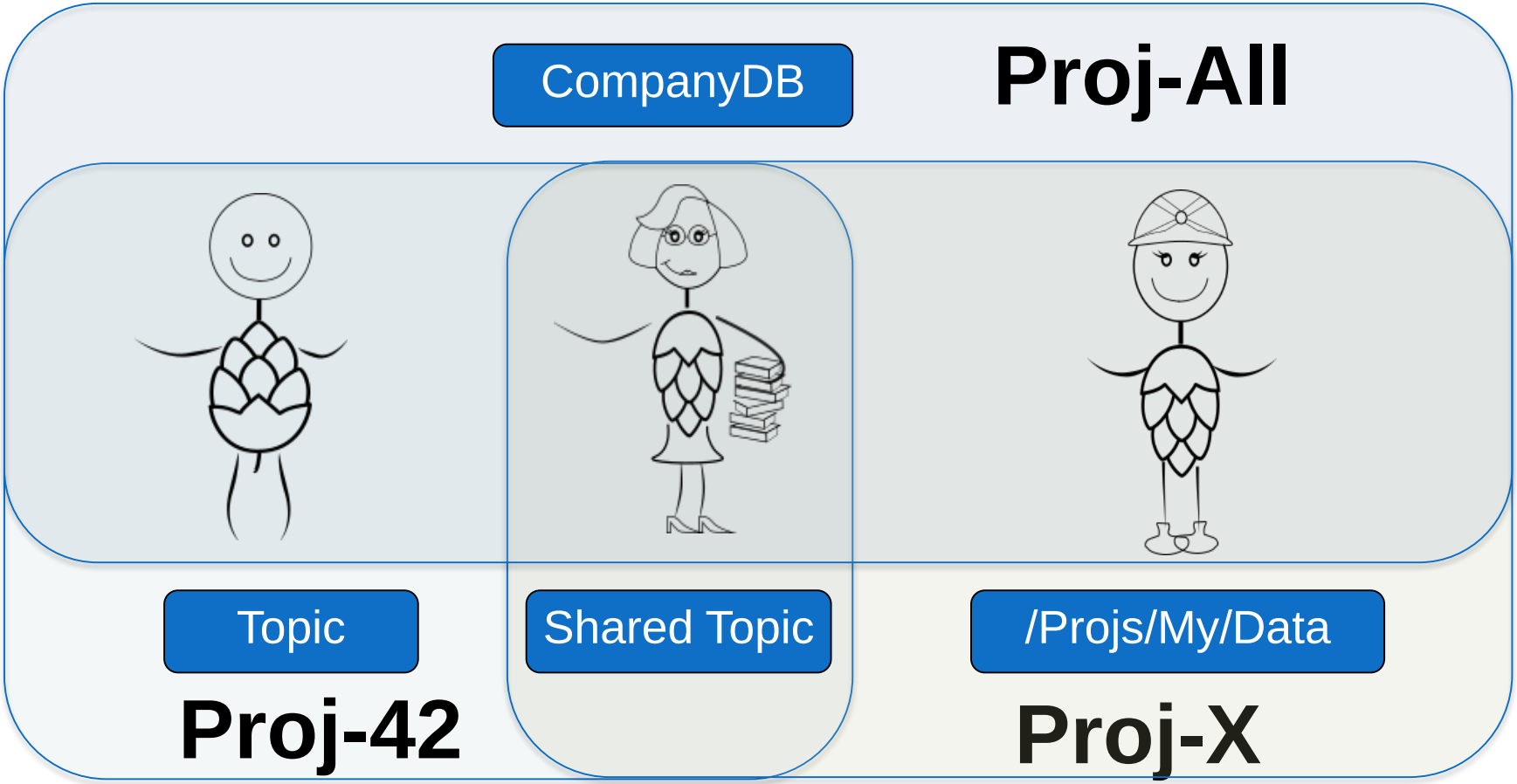
DataSet X

Concepts in Hops Hadoop

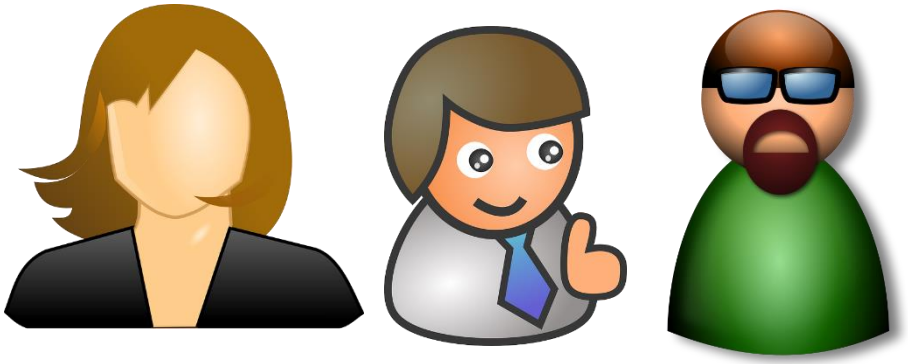
- Concepts:
  - **Facebook:** Friend, Post, Message, Event, Page, Group.
  - **Slack:** Team, Member, Channel, Message, Reaction, Thread.
  - **Hadoop:** Clusters, Users, Apps, Jobs, Files, ACLs/Policies, Kerberos
  - **Databricks:** Clusters, Users , Jobs, Notebooks

- **Hops**: Projects, Datasets, Users, Jobs, Notebooks

Datasets and Projects

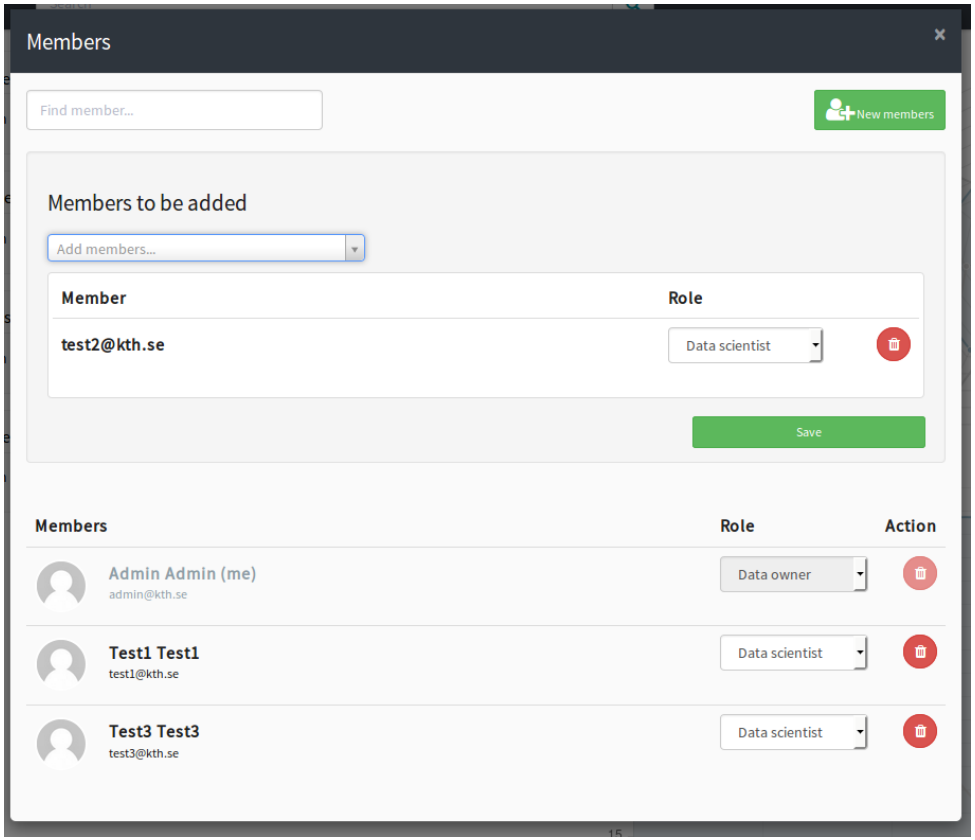


Project



- Members
  - Roles: Owner, Data Scientist
- DataSets
  - Home project
  - Can be shared

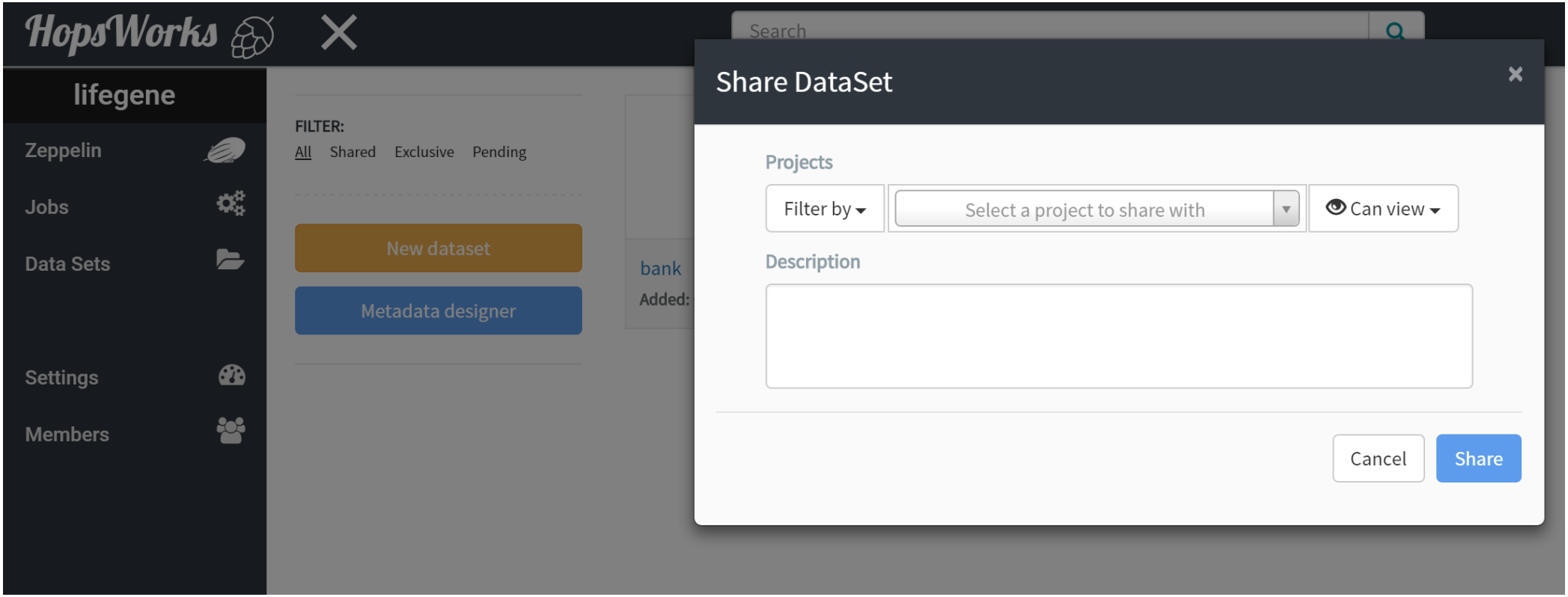
Self-Service Data Access



- Data Owner Privileges
  - Import/Export data
  - Manage Membership
  - Share DataSets
- Data Scientist Privileges
  - Write/Run code
  - Write to StickyBit Datasets
  - Request access to DataSets

We delegate administration of privileges to users

Sharing DataSets between Projects



The same as Sharing Folders in Dropbox

Today's Lab

- Work with Spark, DataFrames, HDFS, and Parquet
- Use Jupyter Notebook with SparkMagic Interpreter
- Introduce the Spark UI to debug performance bugs in Jobs
- Work with Adam/Spark
- Reference tutorials:

https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html

https://www.analyticsvidhya.com/blog/2016/10/spark-dataframe-and-operations/

https://github.com/jupyter-incubator/sparkmagic/blob/master/examples/Pyspark%20Kernel.ipynb

Tasks

1. Register account:
   http://snurran.sics.se:9191/hopsworks

2. Take a Spark Tour

3. Create a Project
   1. Import the shared dataset ‘genomics’
   2. Create a dataset in your project
   3. Add a ‘friend’ to your project as a ‘Data Scientist’
   4. Create a Jupyter Notebook

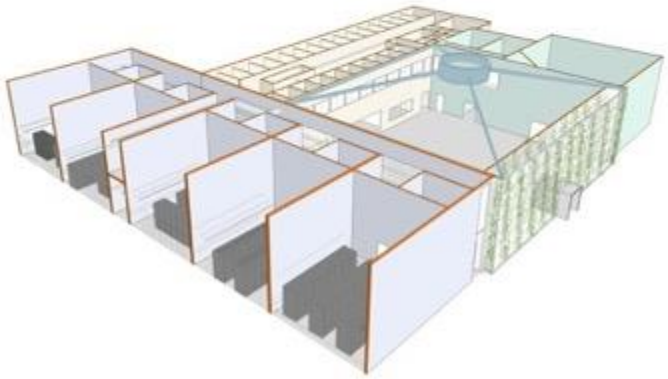
4. Databricks PySpark Tutorial

5. Adam/Scala/Spark Tutorial

SICS ICE research facility

A datacenter research and test environment

Purpose: Increase knowledge, strengthen universities, companies and researchers



SWEDISH ICT SICS

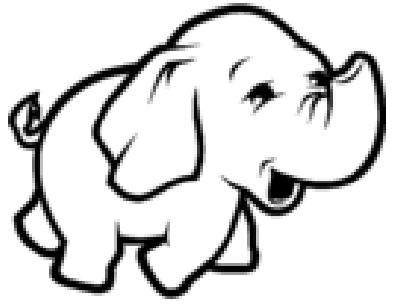
LULEÅ TEKNISKA UNIVERSITET

R&D institute, 2 lab modules, 3-400 servers, 2-3000 square meters

A Data Center Optimized for Data Science

First 140 Dell servers in module 1

- 3600 cores
- 40 TB RAM
- Up to 7.5 petabyte storage
- 10/40 Gb/s network
- Separate management network



Conclusions

- Hops, a new platform for Data Science
  - Spark, Flink, Tensorflow
  - HopsFS
  - Anaconda
  - Extended MetaData

- Support for genomics platforms, like ADAM

The Hops Team

Active:
Jim Dowling, Seif Haridi, Tor Björn Minde, Gautier Berthou, Salman Niazi, Mahmoud Ismail, Theofilos Kakantousis, Ermias Gebremeskel, Antonios Kouzoupis, Alex Ormenisan, Roberto Bampi, Fabio Buso, Fanti Machmount Al Samisti, Braulio Grana, Adam Alpire, Zahin Azher Rashid, Robin Andersso, ArunaKumari Yedurupaka, Tobias Johansson, August Bonds, Filotas Siskos.

www.hops.io

@hopshadoop

Alumni:

Vasileios Giannokostas, Johan Svedlund Nordström,Rizvi Hasan, Paul Mälzer, Bram Leenders, Juan Roca, Misganu Dessalegn, K “Sri” Srijeyanthan, Jude D’Souza, Alberto Lorente, Andre Moré, Ali Gholami, Davis Jaunzems, Stig Viaene, Hooman Peiro, Evangelos Savvidis, Steffen Grohsschmiedt, Qi Qi, Gayana Chandrasekara, Nikolaos Stanogias, Daniel Bali, Ioannis Kerkinos, Peter Buechler, Pushparaj Motamari, Hamid Afzali, Wasif Malik, Lalith Suresh, Mariano Valles, Ying Lieu.

RI. SE

eit Digital

Stiftelsen för Strategisk Forskning

KTH Vetenskap och Konst

European Commission

Logical Clocks

Ericsson

Oracle

Karolinska Institutet

Scania

yanzi networks

MTA SZTAKI

Atos