RHEEM

A Cross-Platform Big Data system

Turning a Zoo into a Circus…

Jorge Quiané

Data Analytics @ QCRI

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

مؤسسة قطر
Qatar Foundation

Big Data is like a Zoo!

Go fast?

Endurance?

Go to the forest?

Climb the everest?

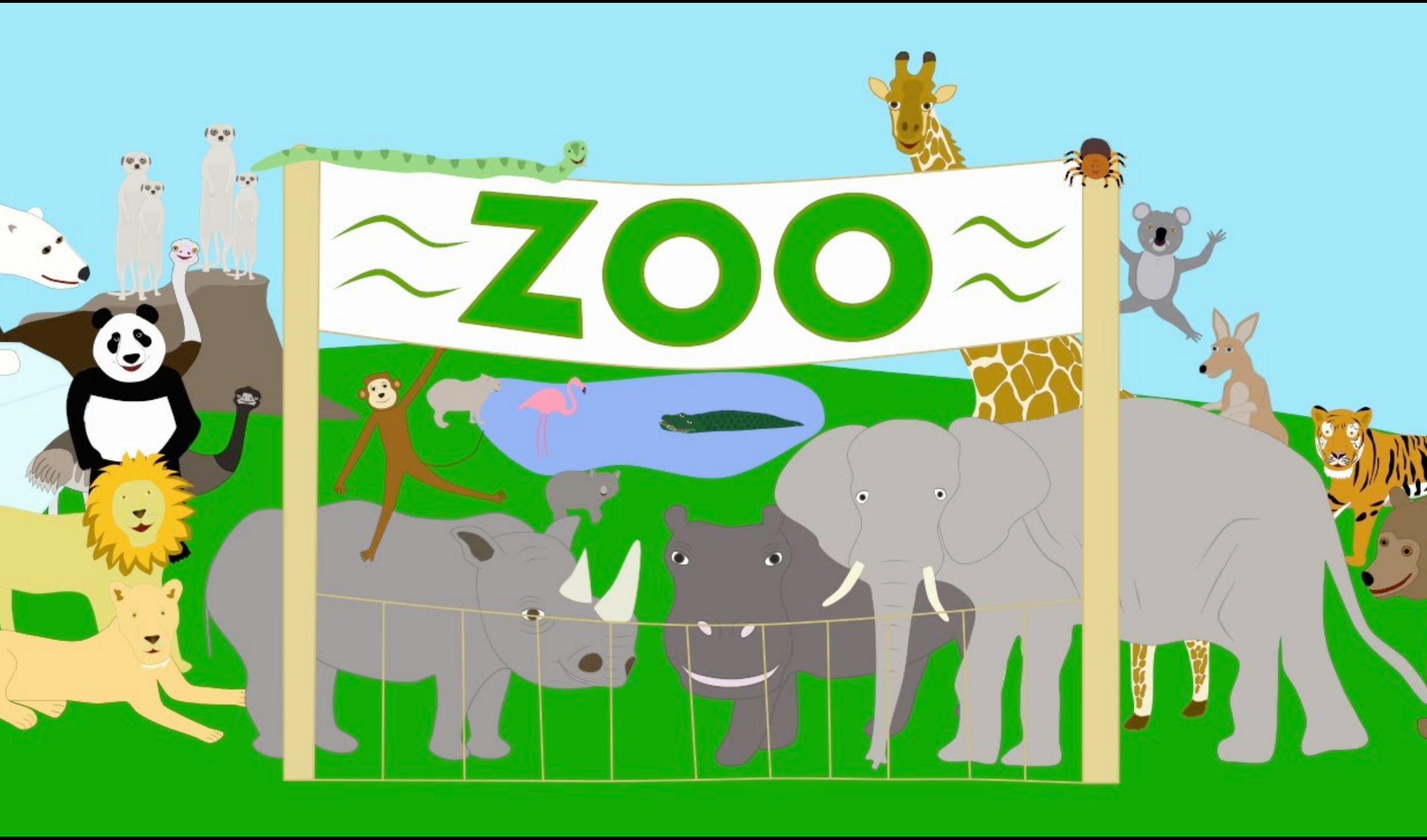
Want to fly?

Fly!

One Animal

Does Not

Fit All

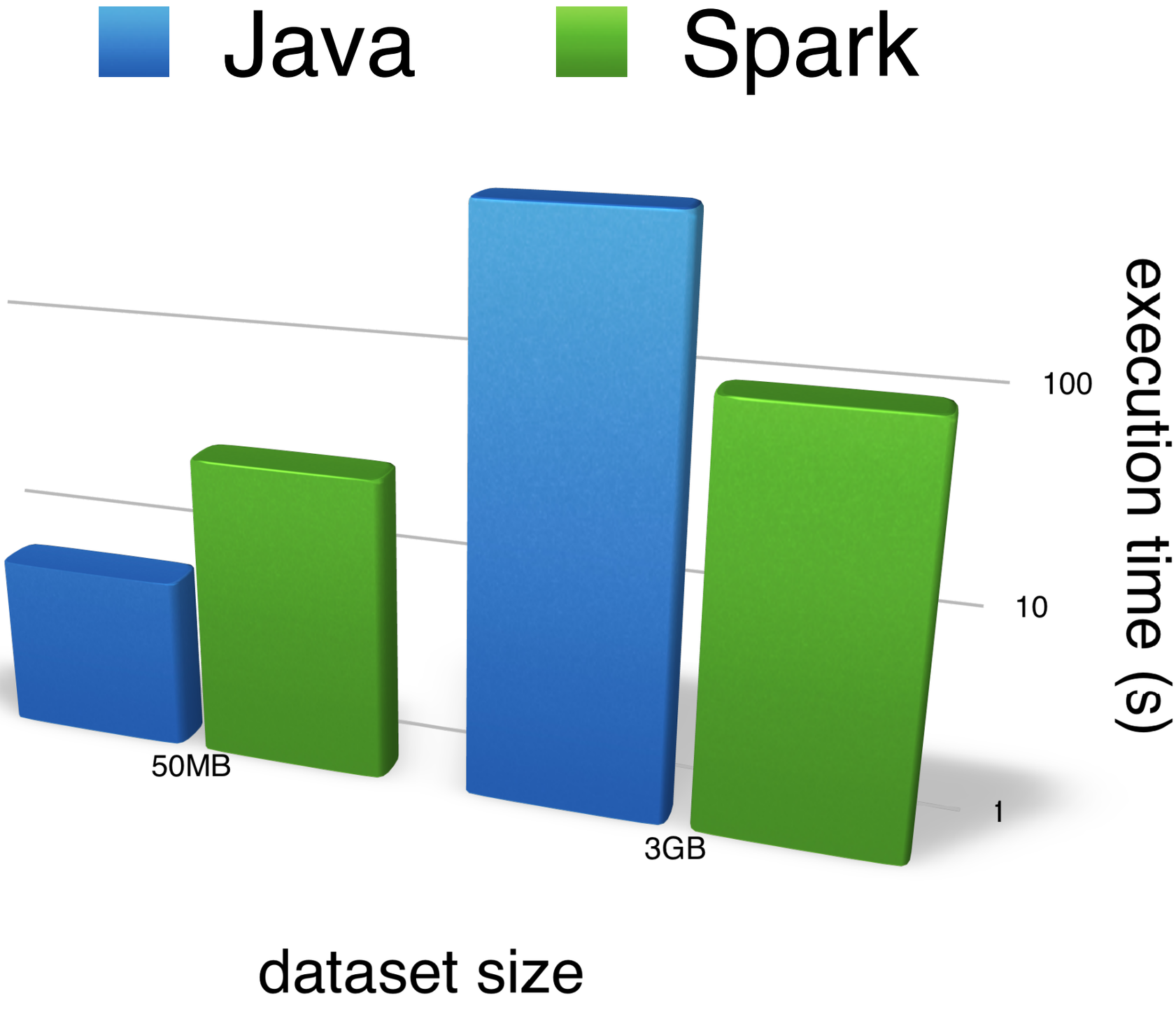

ZOO

One Size

Does Not

Fit All

Google Index



Big Data Landscape

Big Data Landscape

NoSQL Databases

NewSQL Databases

BI Platforms

Statistical Computing

Log Analytics

Social Analytics

Ad Optimization

Security

Vertical AI Applications

Graph Databases

MPP Databases

Cloud EDW

Data Transformation

Data Integration

Real-Time

Machine Learning

Speech & NLP

Horizontal AI

Publisher Tools

Govt/ Regulation

Finance

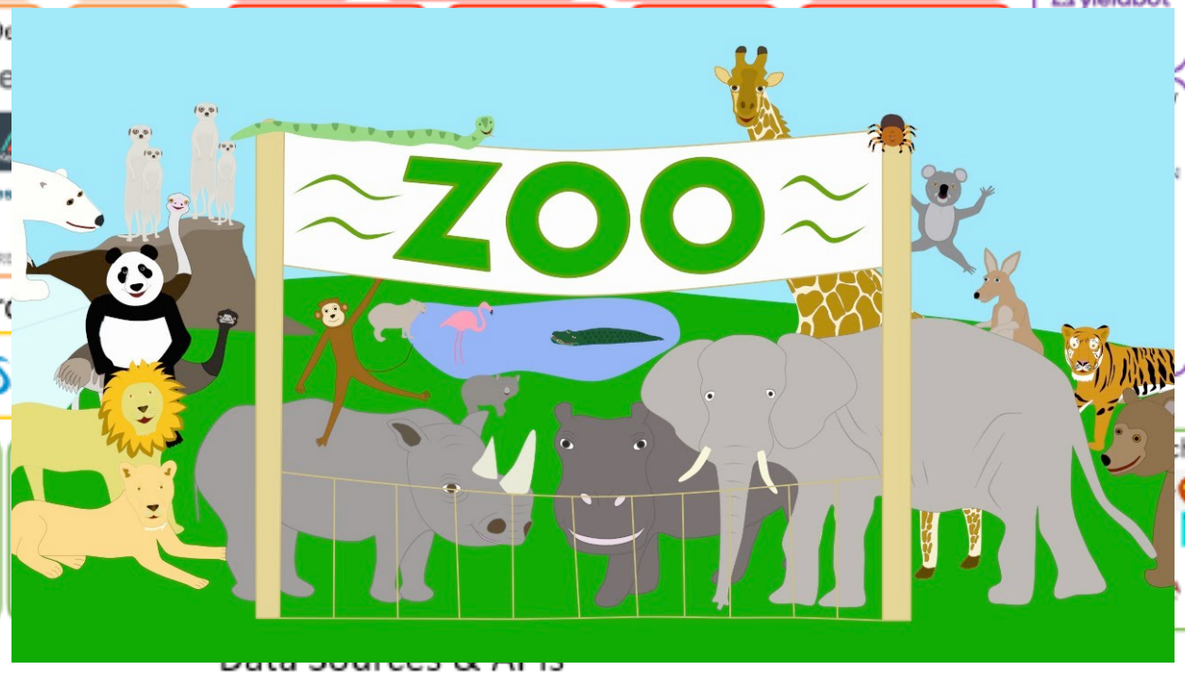
Management / Monitoring

Security

Storage

Life Sciences

Industries



Open Source

Framework

Query / Data Flow

Machine Learning

Search

Security

Visualization

Data Sources & APIs

Health

IOT

Financial & Economic Data

Air / Space / Sea

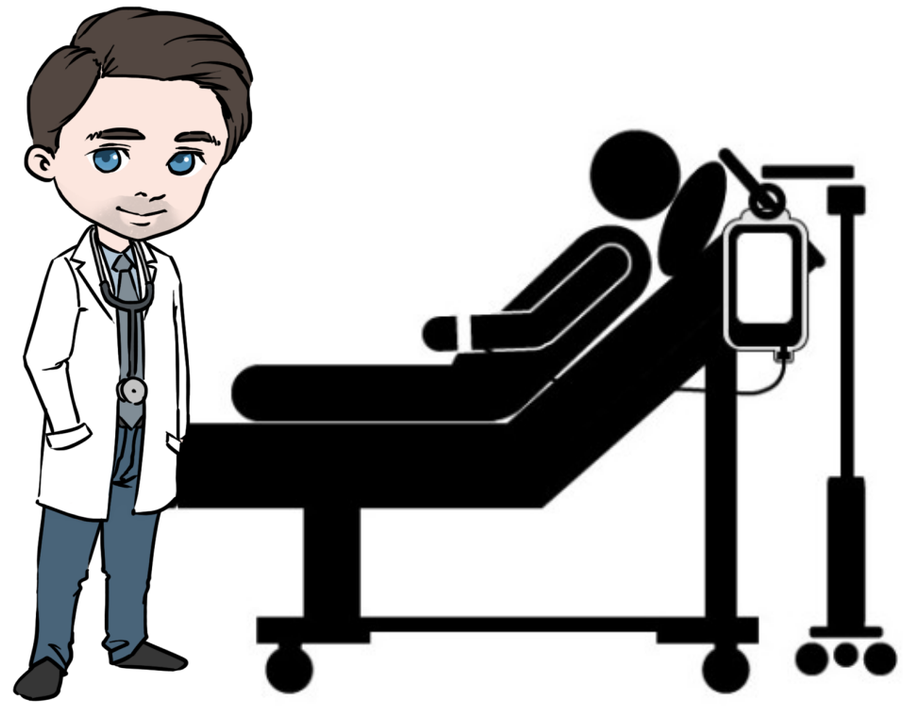
Location/People/Entities

Other

Incubators & Schools

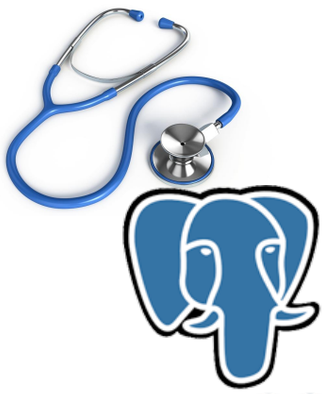
Healthcare

— A Big Data Revolution —



Medical records

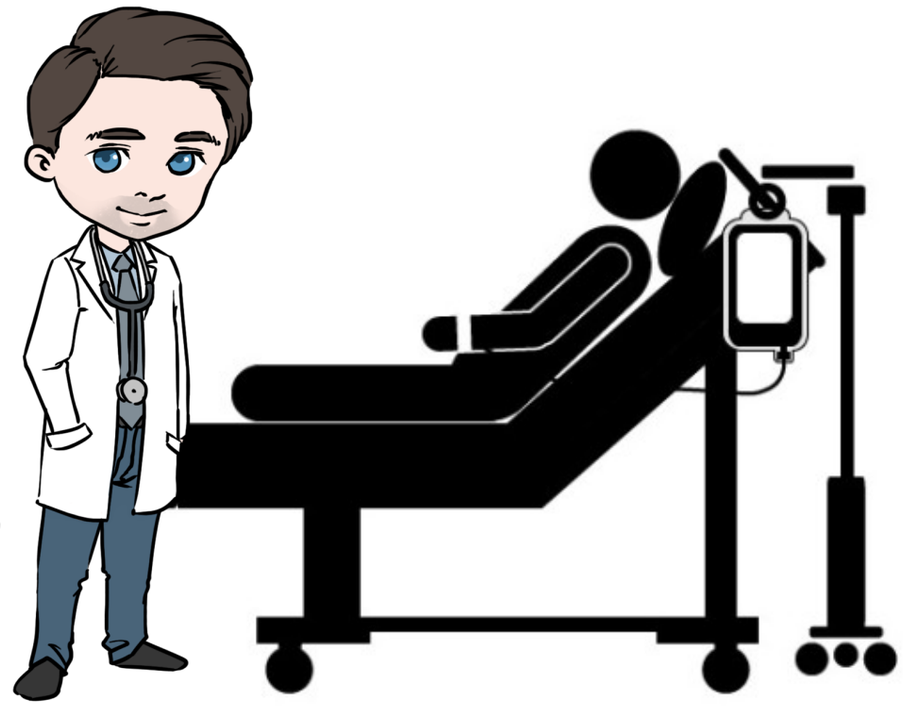
Medical devices



PostgreSQL RethinkDB

Healthcare

— A Big Data Revolution —



Patient current health

Medical records

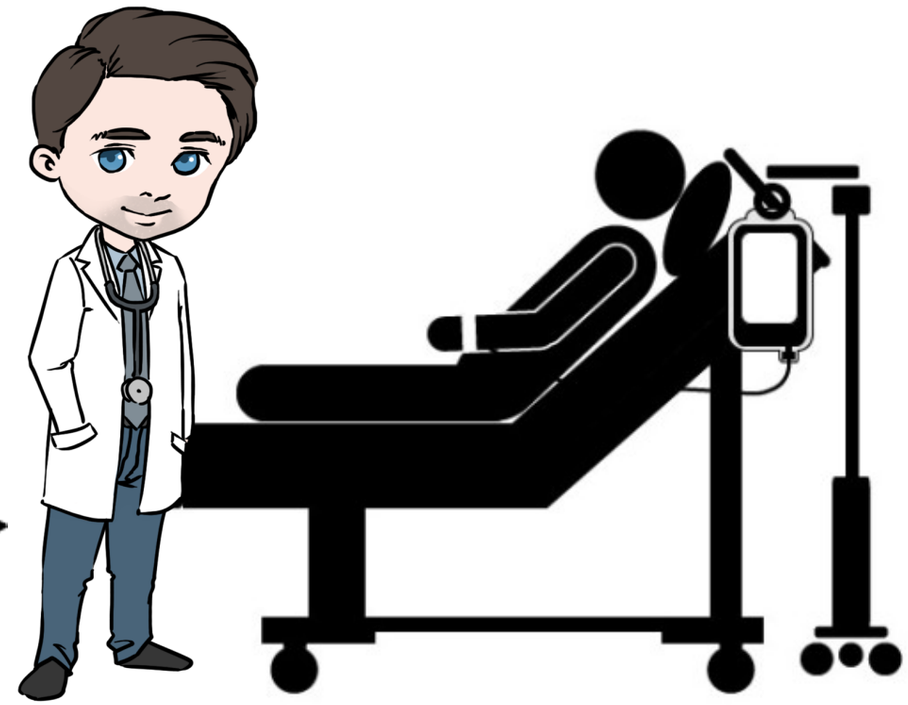
Medical devices

PostgreSQL

RethinkDB

Healthcare

— A Big Data Revolution —

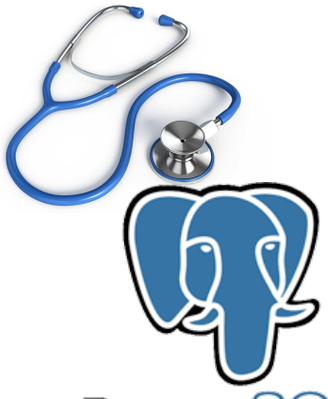


Patient current health

Patient's records

Medical records

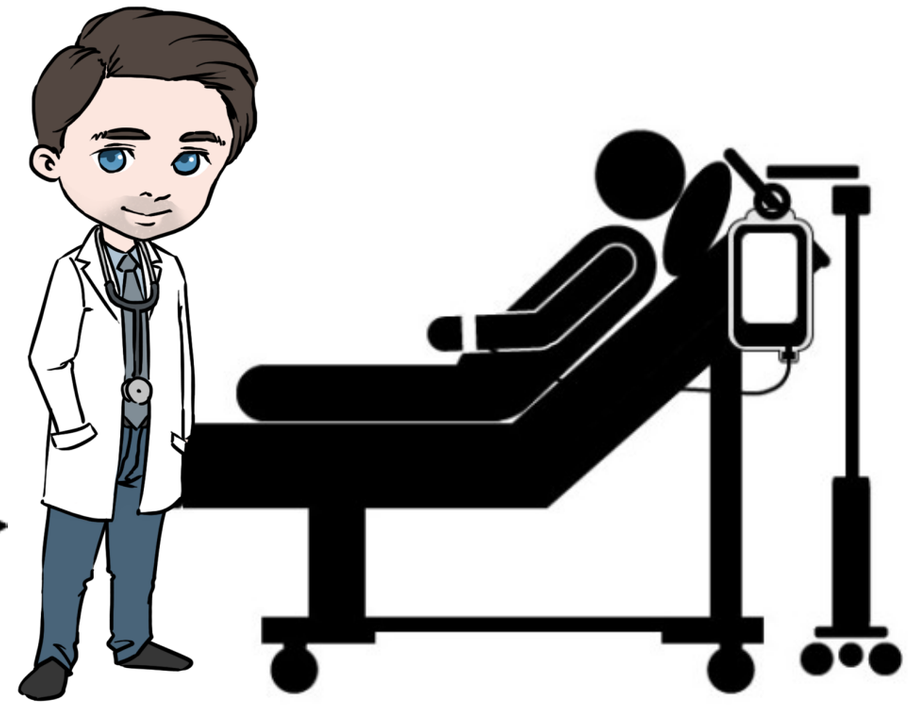
Medical devices



PostgreSQL RethinkDB

Healthcare

— A Big Data Revolution —



Patient current health

Patient's records

Medical records

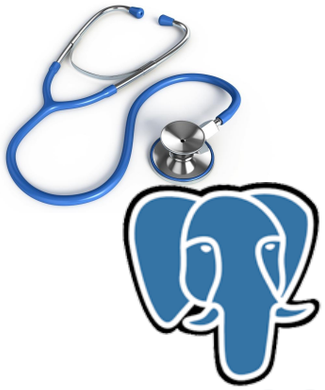
Medical devices

Insurance files

Genome data

Wearable devices

Social data

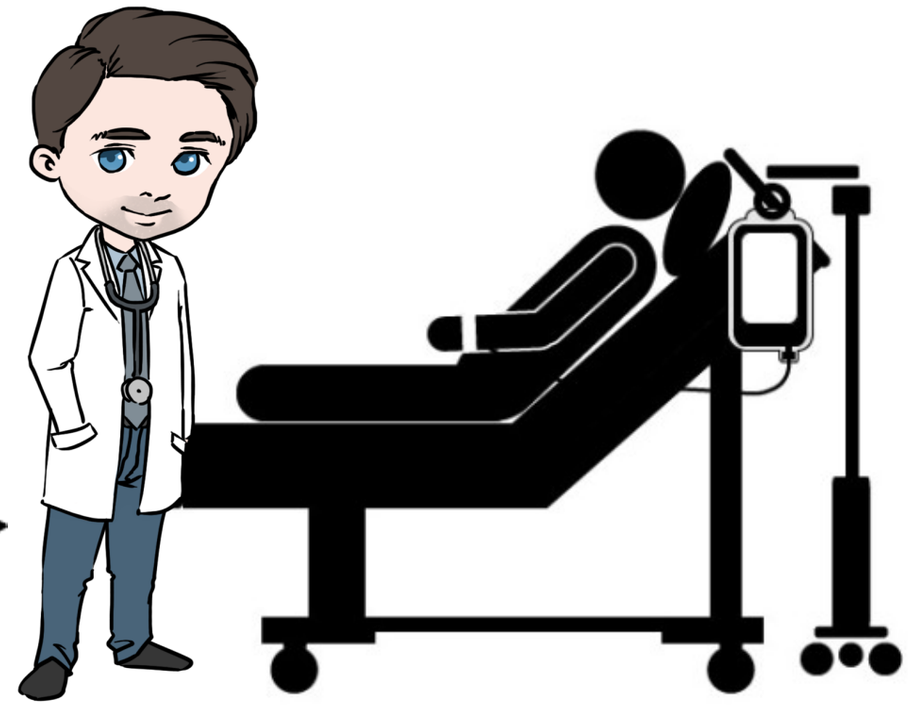


PostgreSQL

RethinkDB

Healthcare

— A Big Data Revolution —



Records of all similar world-wide patients

Patient current health

Patient records

Medical records

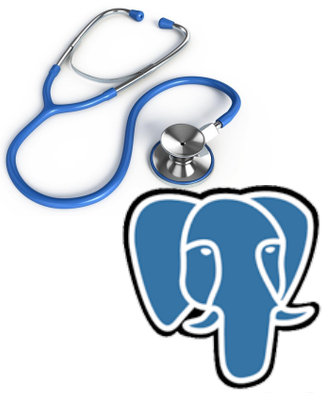
Medical devices

Insurance files

Genome data

Wearable devices

Social data

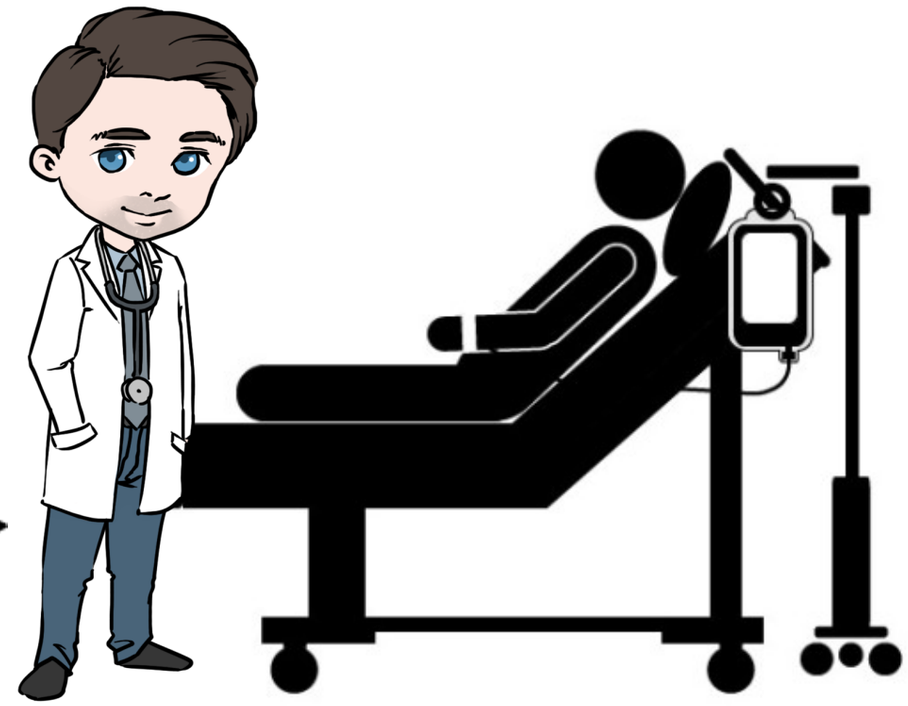


PostgreSQL RethinkDB

mongoDB

Healthcare

— A Big Data Revolution —



Records of all similar world-wide patients

Patient current health

Patient records

Records of patient's medication

Medical records

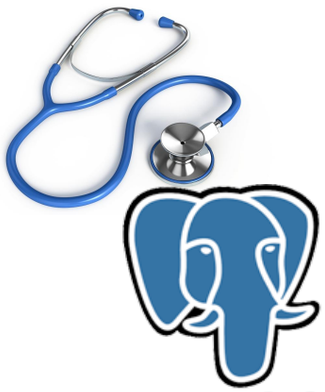
Medical devices

Insurance files

Genome data

Wearable devices

Social data

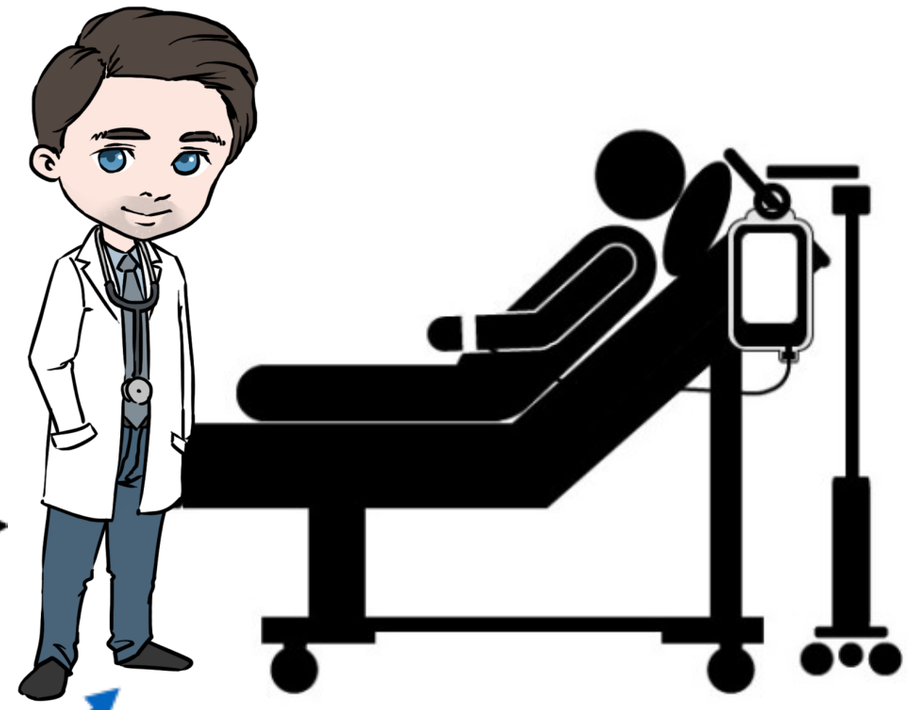


PostgreSQL RethinkDB

mongoDB

Healthcare

— A Big Data Revolution —



Records of all similar world-wide patients

Patient current health

Patient records

Records of patient's medication

Patient's DNA

Medical records

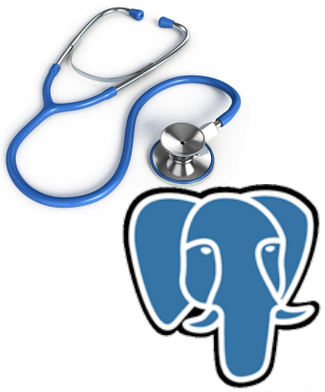
Medical devices

Insurance files

Genome data

Wearable devices

Social data

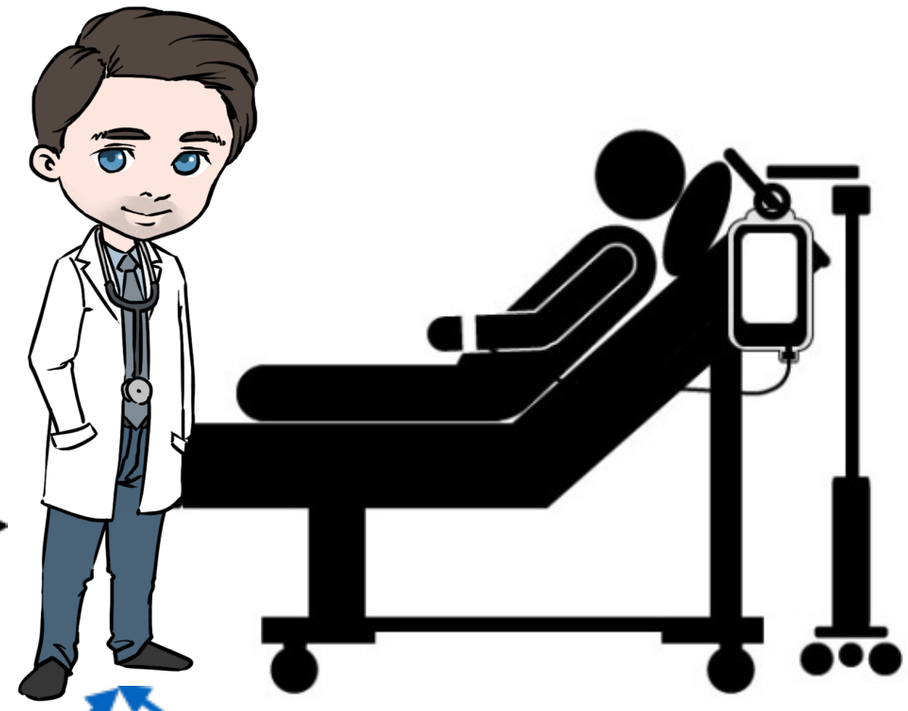


PostgreSQL

RethinkDB

Healthcare

— A Big Data Revolution —



Records of all similar world-wide patients

Patient current health

Patient's records

Records of patient's medication

Patient's DNA

Patient's routine

Medical records

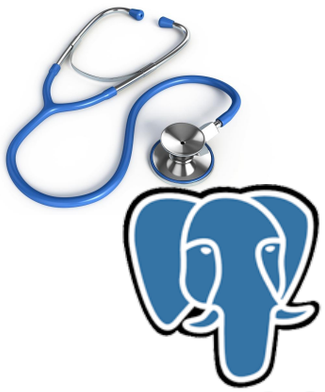
Medical devices

Insurance files

Genome data

Wearable devices

Social data

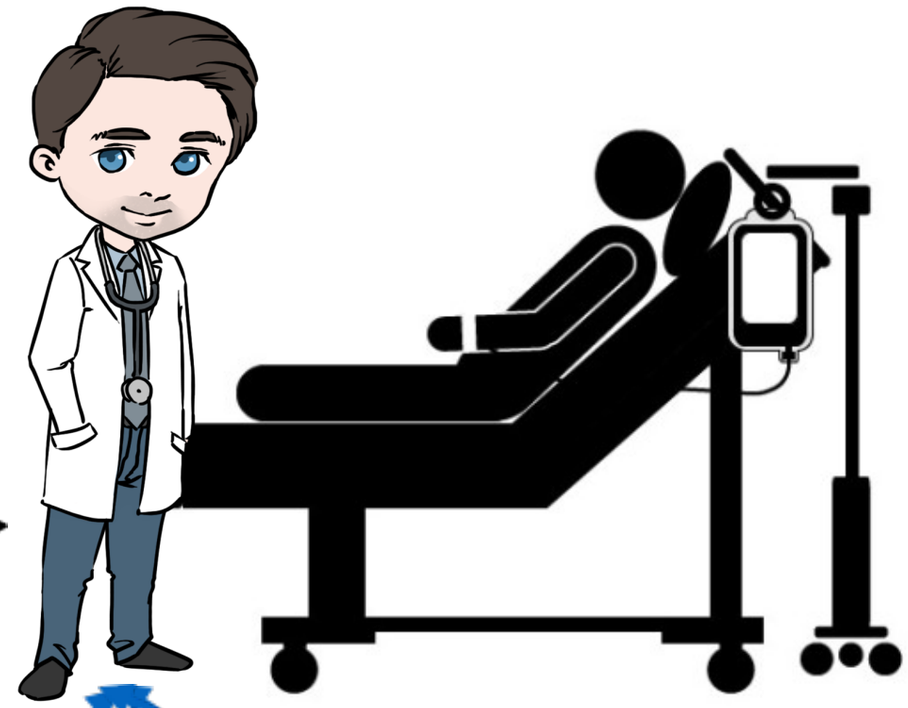


PostgreSQL RethinkDB

mongoDB

Healthcare

— A Big Data Revolution —



Records of all similar world-wide patients

Patient current health

Patient's home environment

Patient's records

Records of patient's medication

Patient's DNA

Patient's habits

Patient's routine

Medical records

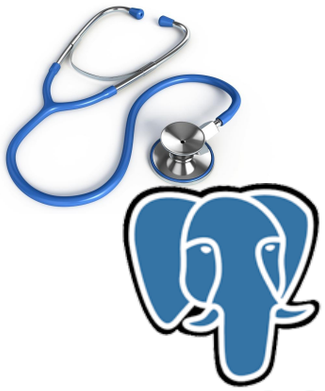
Medical devices

Insurance files

Genome data

Wearable devices

Social data

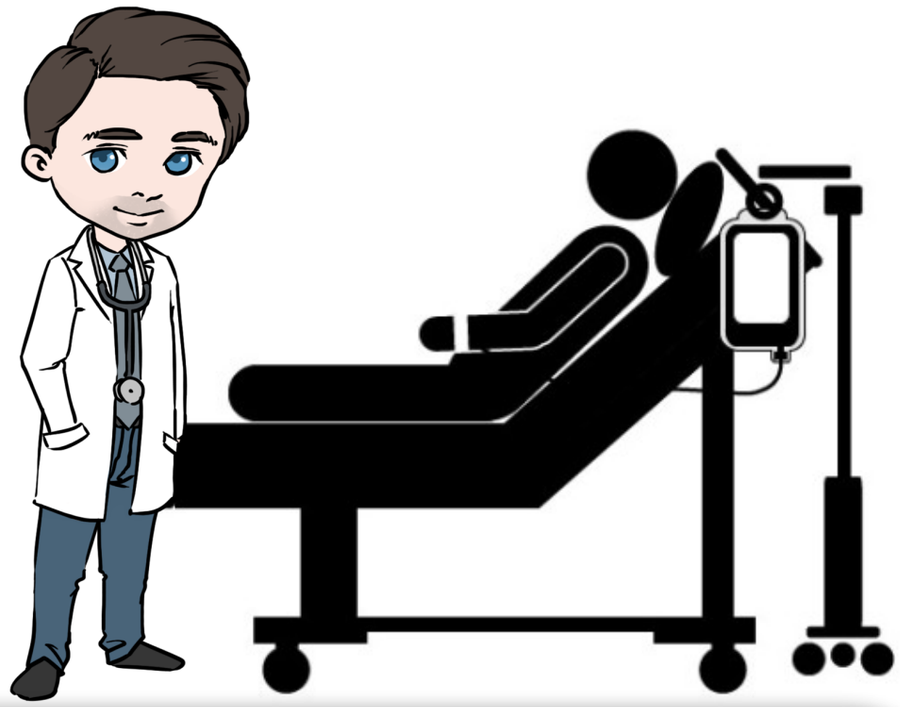


PostgreSQL RethinkDB

mongoDB

Healthcare

— A Big Data Revolution —



Big Data Framework

Data Analytics

Records of all similar world-wide patients

Patient current health

Patient's records

Records of patient's medication

Patient's DNA

Patient's routine

Patient's home environment

Patient's habits

Medical records

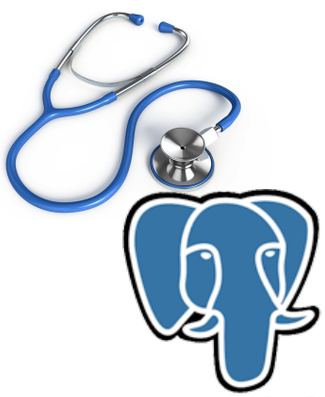
Medical devices

Insurance files

Genome data

Wearable devices

Social data



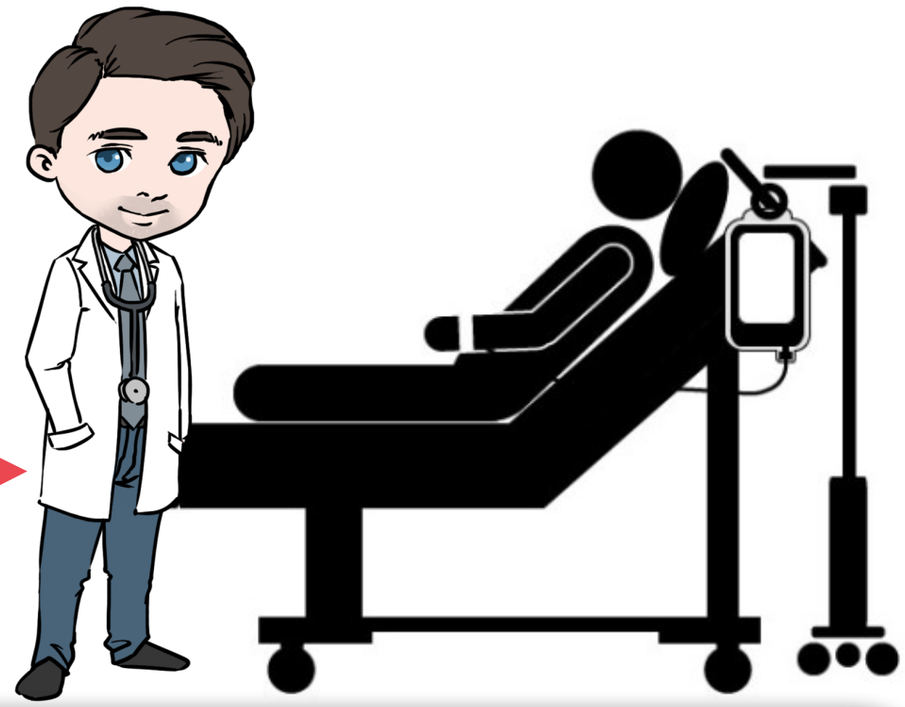
PostgreSQL

RethinkDB

mongoDB

Healthcare

— A Big Data Revolution —



diagnosis & prescription

Big Data Framework

Data Analytics

Records of all similar world-wide patients

Patient current health

Records of patient's medication

Patient's DNA

Patient's routine

Patient's home environment

Patient's habits

Patient's records

Medical records

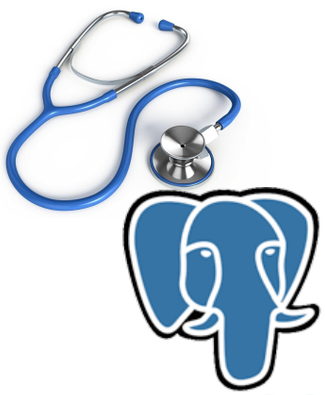
Medical devices

Insurance files

Genome data

Wearable devices

Social data



PostgreSQL

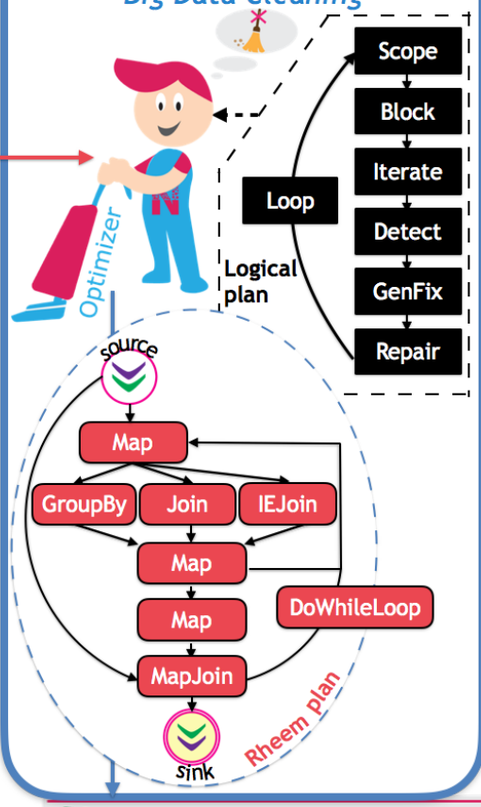
RethinkDB

RHEEM Apps

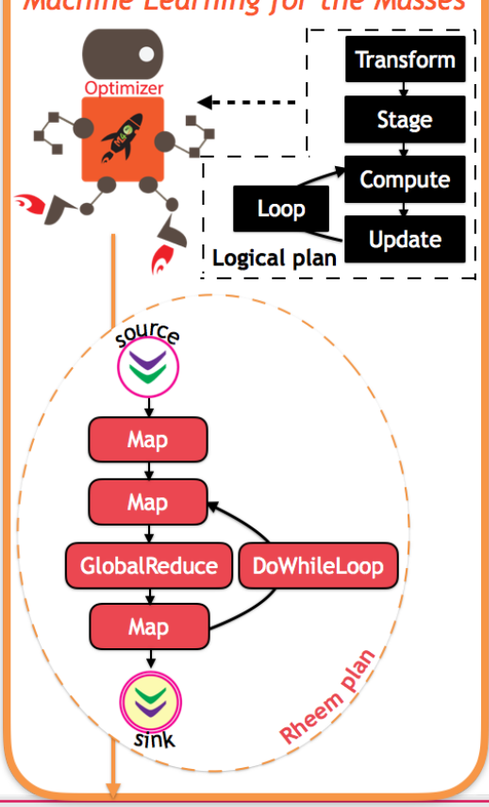
Algorithmic Optimizations

Logical Operators

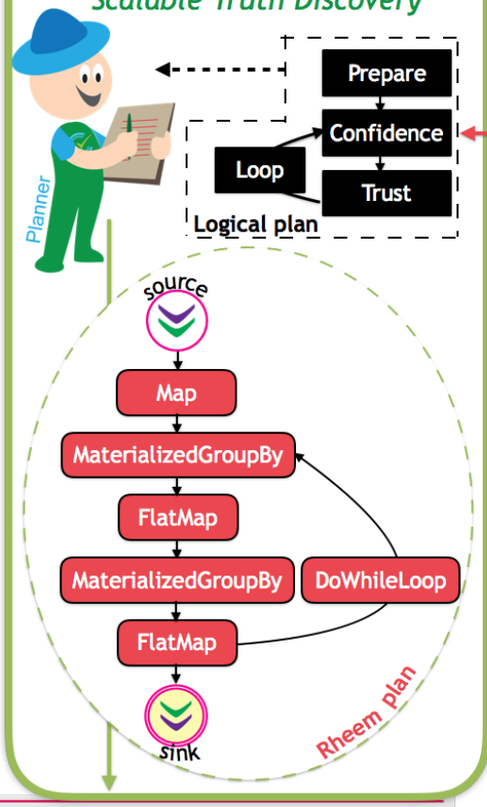
Nadeef
Big Data Cleaning



ML4all
Machine Learning for the Masses



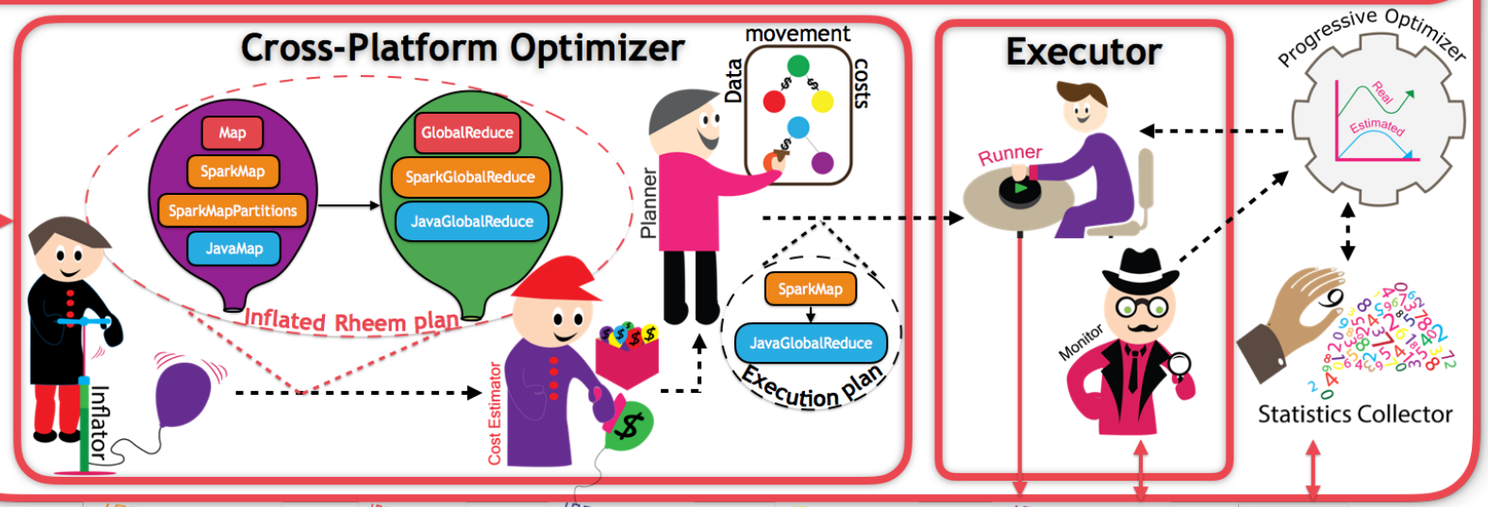
DAFNA
Scalable Truth Discovery



RHEEM Operators

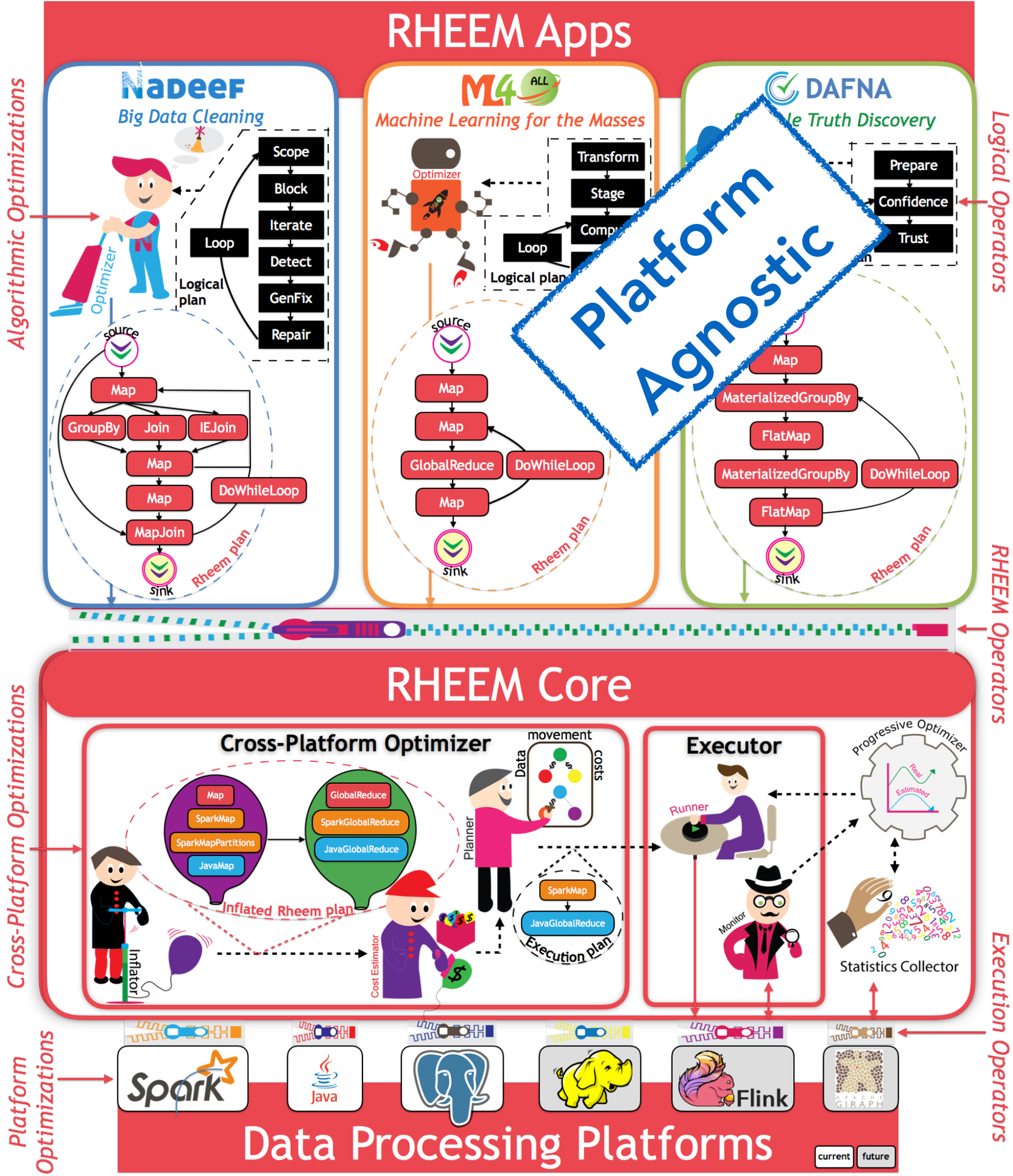
RHEEM Core

Cross-Platform Optimizations



Platform Optimizations

Execution Operators



RHEEM Apps
Algorithmic Optimizations
Logical Operators
NADEEF
Big Data Cleaning
Scope
Block
Iterate
Detect
GenFix
Repair
Loop
Logical plan
Optimizer
source
Map
GroupBy
Join
IEJoin
Map
Map
MapJoin
DoWhileLoop
sink
Rheem plan
ML4 ALL
Machine Learning for the Masses
Optimizer
Transform
Stage
Loop
Logical plan
source
Map
Map
GlobalReduce
DoWhileLoop
Map
sink
Rheem plan
DAFNA
Truth Discovery
Prepare
Confidence
Trust
Map
MaterializedGroupBy
FlatMap
MaterializedGroupBy
DoWhileLoop
FlatMap
sink
Rheem plan
Platform Agnostic
RHEEM Operators
RHEEM Core
Cross-Platform Optimizations
Cross-Platform Optimizer
Map
SparkMap
SparkMapPartitions
JavaMap
GlobalReduce
SparkGlobalReduce
JavaGlobalReduce
Inflated Rheem plan
Inflator
Cost Estimator
Planner
Data movement costs
SparkMap
JavaGlobalReduce
Execution plan
Executor
Runner
Monitor
Progressive Optimizer
Real
Estimated
Statistics Collector
Execution Operators
Platform Optimizations
Spark
Java
Flink
APACHE GIRAPH
Data Processing Platforms
current
future

RHEEM Apps

Algorithmic Optimizations

Logical Operators

RHEEM Operators

Cross-Platform Optimizations

Execution Operators

Platform Optimizations

Platform Agnostic

What is the best platform?

Nadeef
Big Data Cleaning

Scope
Block
Iterate
Detect
GenFix
Repair
Loop
Logical plan
Optimizer
source
Map
GroupBy
Join
IEJoin
Map
Map
MapJoin
DoWhileLoop
sink
Rheem plan

ML4ALL
Machine Learning for the Masses

Transform
Stage
Loop
Logical plan
Optimizer
source
Map
Map
GlobalReduce
DoWhileLoop
Map
sink
Rheem plan

DAFNA
Truth Discovery

Prepare
Confidence
Trust
Map
MaterializedGroupBy
FlatMap
MaterializedGroupBy
DoWhileLoop
FlatMap
sink
Rheem plan

RHEEM Core

Optimizer
movement
Data
costs
Planner
SparkGlobalReduce
JavaGlobalReduce
Inflated Rheem plan
Cost Estimator
SparkMap
JavaGlobalReduce
Execution plan

Executor
Runner
Monitor
Progressive Optimizer
Real
Estimated
Statistics Collector

Spark
Java
Flink
APACHE GIRAPH

Data Processing Platforms

current
future

RHEEM Apps

Algorithmic Optimizations

Logical Operators

RHEEM Operators

Cross-Platform Optimizations

Platform Optimizations

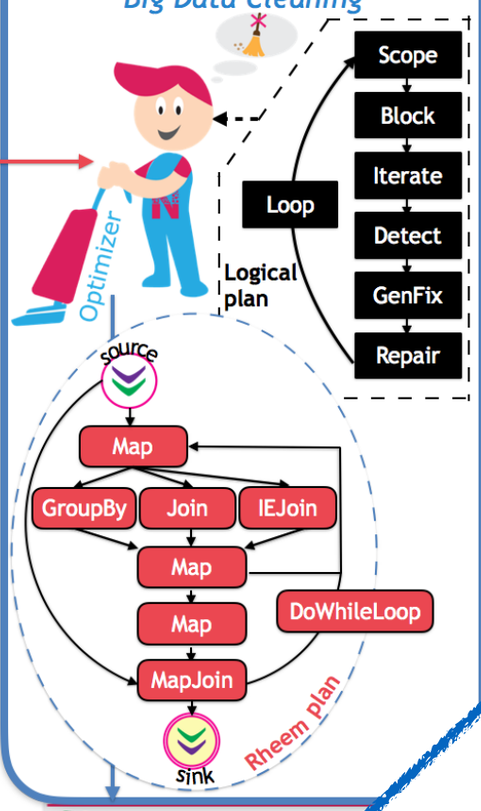
Execution Operators

Platform Agnostic

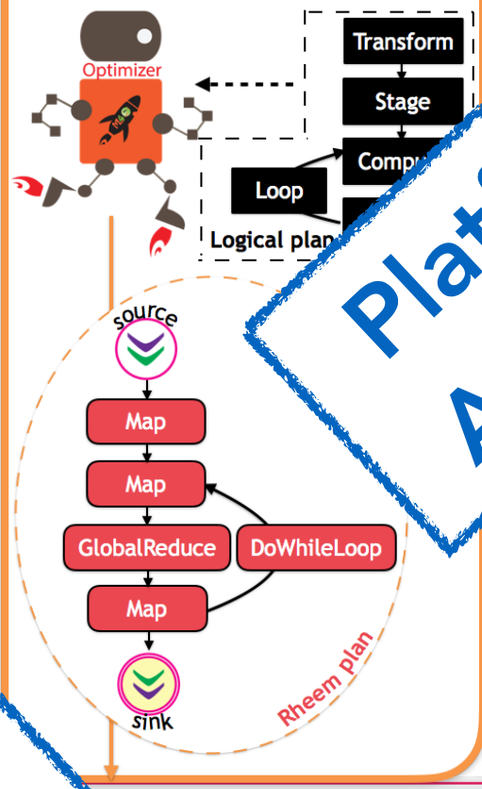
What is the best platform?

Always Evolving

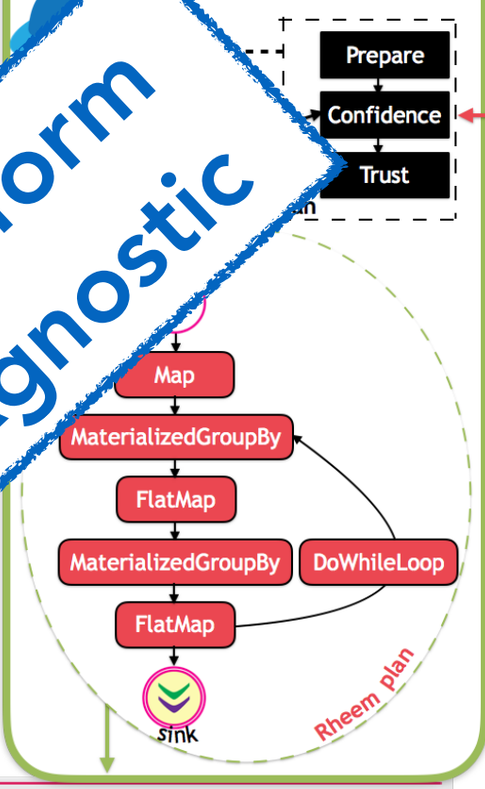
Nadeef
Big Data Cleaning



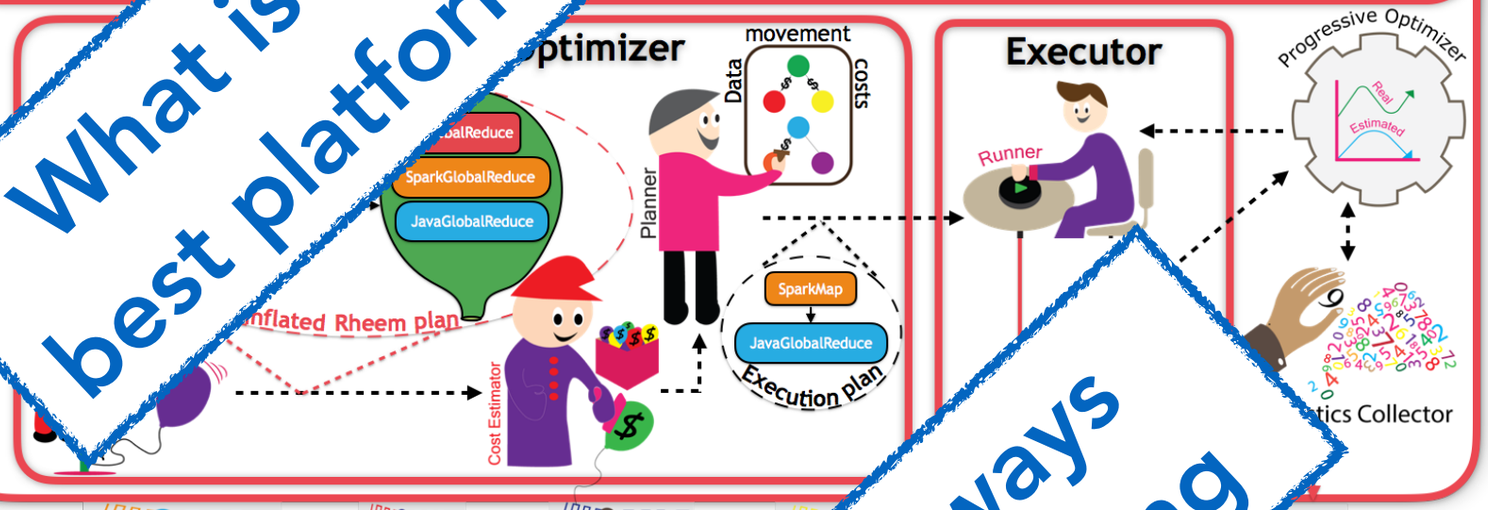
ML4ALL
Machine Learning for the Masses



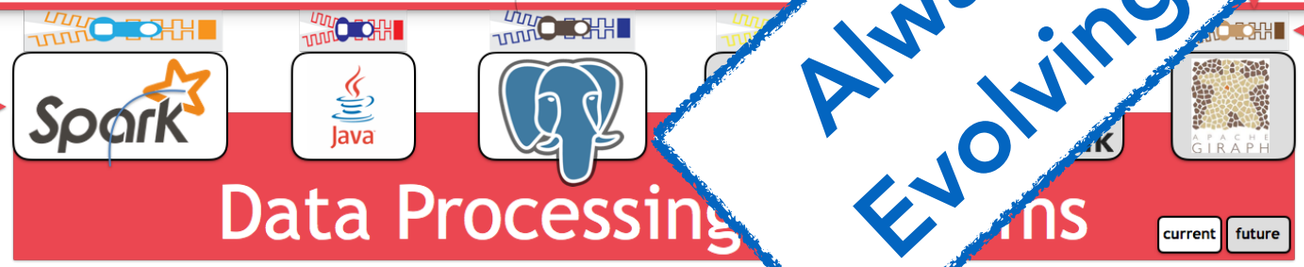
DAFNA
Big Truth Discovery



RHEEM Core



Data Processing



Rheem to the Rescue

Healthcare

Rheem to the Rescue

Healthcare

Smart Cities

Rheem to the Rescue

Healthcare

Smart Cities

Cyber Security

Rheem to the Rescue

Healthcare

Smart Cities

Cyber Security

Sport Performance

Rheem to the Rescue

Healthcare

Smart Cities

Cyber Security

Gas & Oil

Sport Performance

Rheem to the Rescue

Healthcare

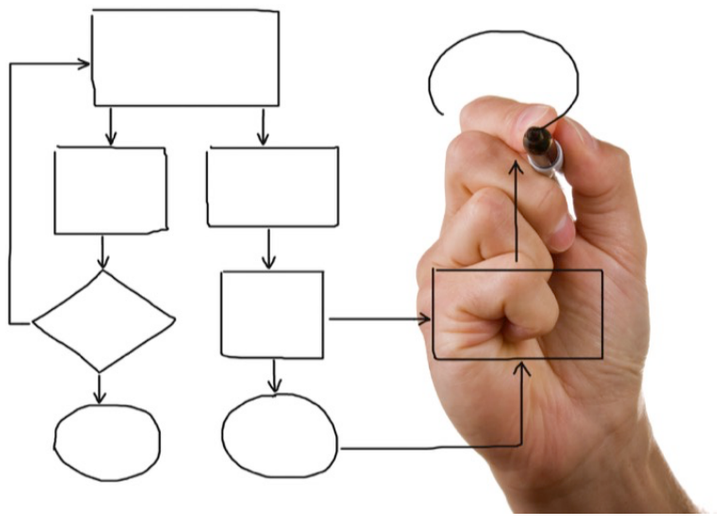
Smart Cities

Cyber Security

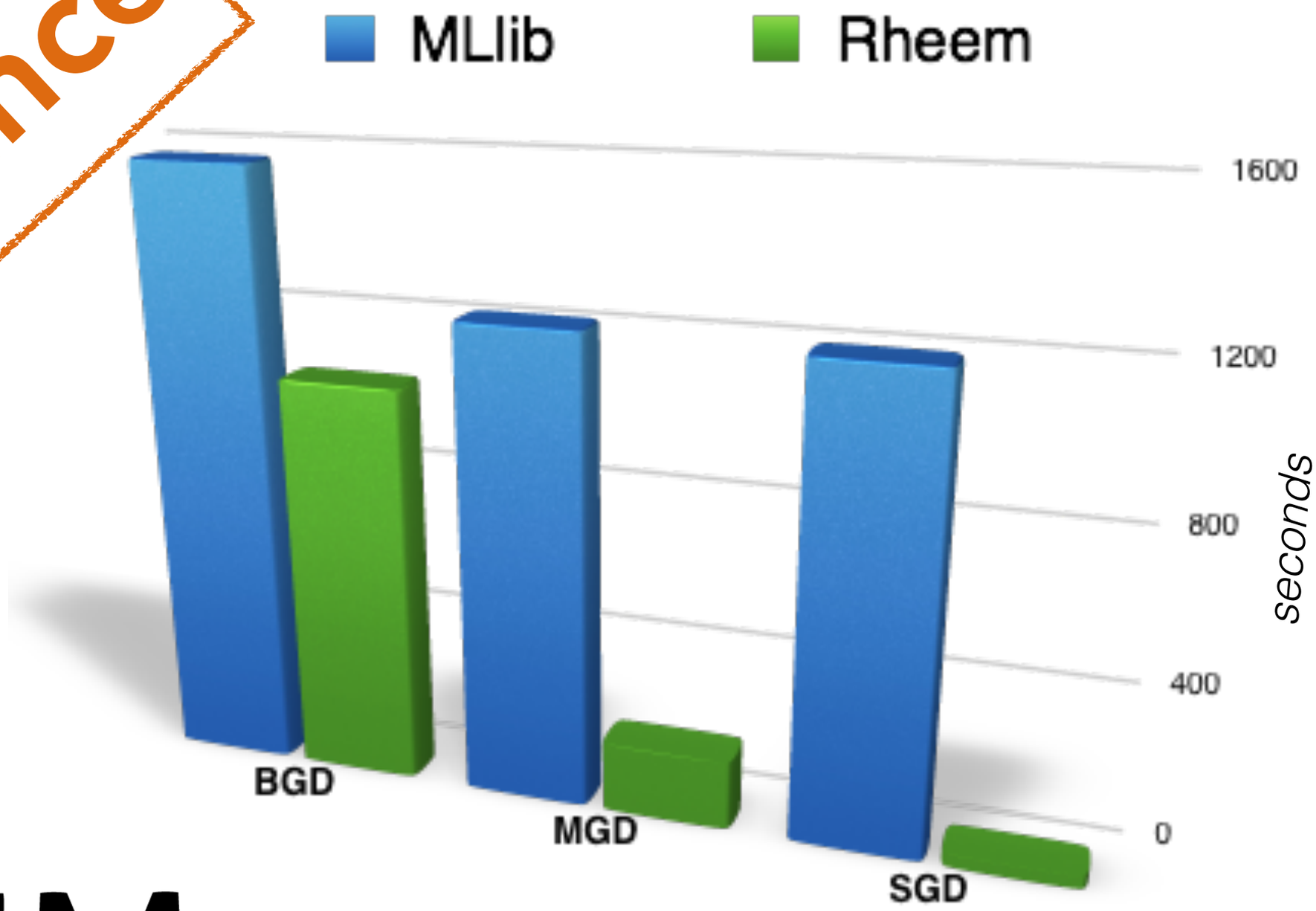
Gas & Oil

Sport Performance

Business Process



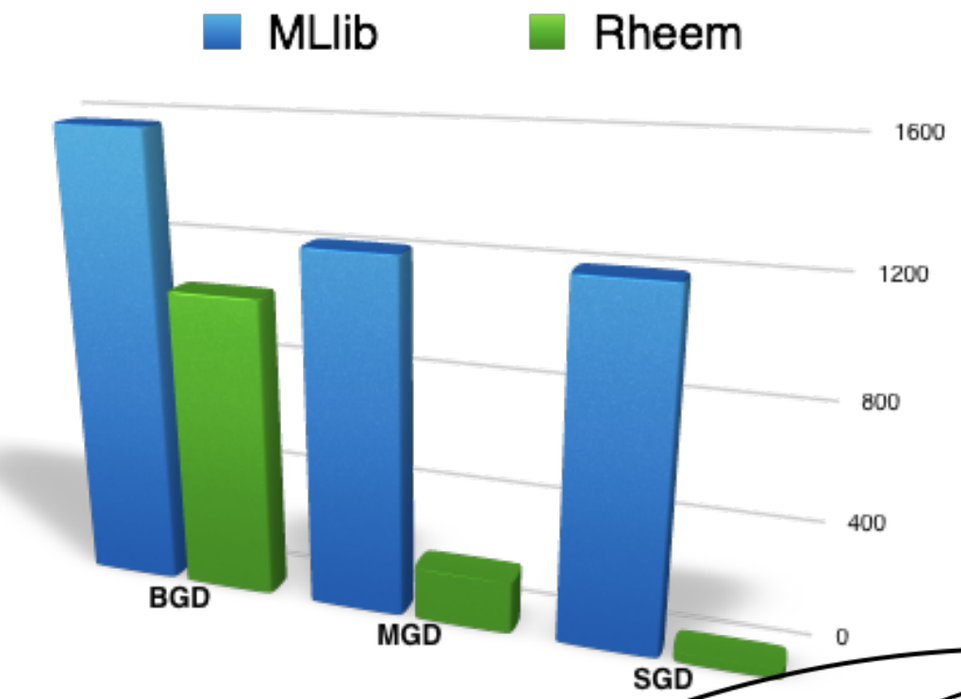
Performance

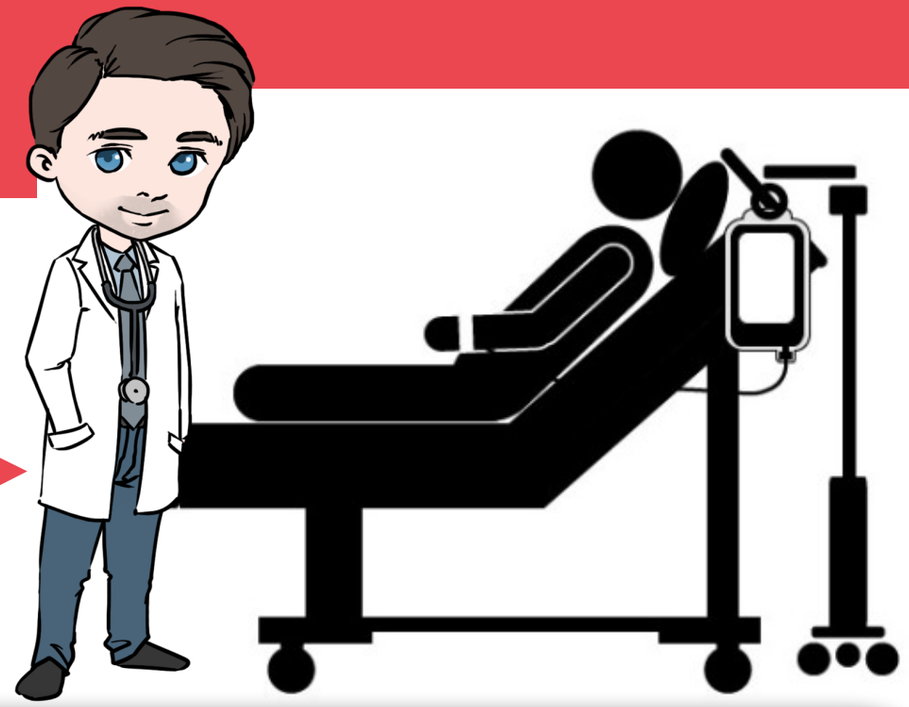


RHEEM

Classification Task

Leveraging Big Data





diagnosis & prescription

RHEEM

Data Analytics

Patient's records

Patient current health

Records of patient's medication

Patient's DNA

Patient's routine

Patient's home environment

Patient's habits

Medical records

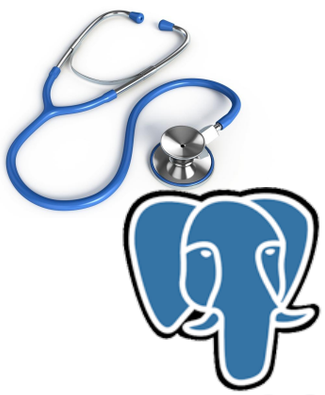
Medical devices

Insurance files

Genome data

Wearable devices

Social data



PostgreSQL

RethinkDB

mongoDB