THE UNIVERSITY OF WARWICK

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

Chiron: Data Intensive Computing for Warwick

TRAINING

PROVIDING RESEARCHERS WITH THE ABILITY TO USE LARGE-SCALE DATA TO UNDERSTAND AND ADDRESS REAL-WORLD CHALLENGES.

THE UNIVERSITY OF WARWICK

Delivering research for real-world problems

What **tools** do we need to **fully exploit** the potential of the **deluge of data** available to researchers today?

- Traffic
- People
- GIS
- Cameras
- Security
- Buildings
- Internet of Things

Delivering research for real-world problems

Large scale analysis comes with its own set of problems, challenges, and opportunities

High Volume, Low Value Data

↓

Low Volume, High Value Information




Delivering research for real-world problems

Large scale analysis comes with its own set of problems, challenges, and opportunities

- Volume
- Velocity
- Variety
- Veracity

- Hardware
- Software
- Skills
- Science

Hardware platforms for scalable analytics

Not wildly different hardware from platforms for HPC (High Performance Computing)

- Scale *out* not *up*
- Network a *cluster* of nodes
- Distribute *tasks* to where *data* is available
- Consider:
  - Communications (Network)
  - Compute (CPUs, Memory, Accelerators)
  - Storage (HDDs / SSDs)

Software platforms for scalable analytics

Two key software paradigms: **Bulk** (offline) and **Streaming** (online)

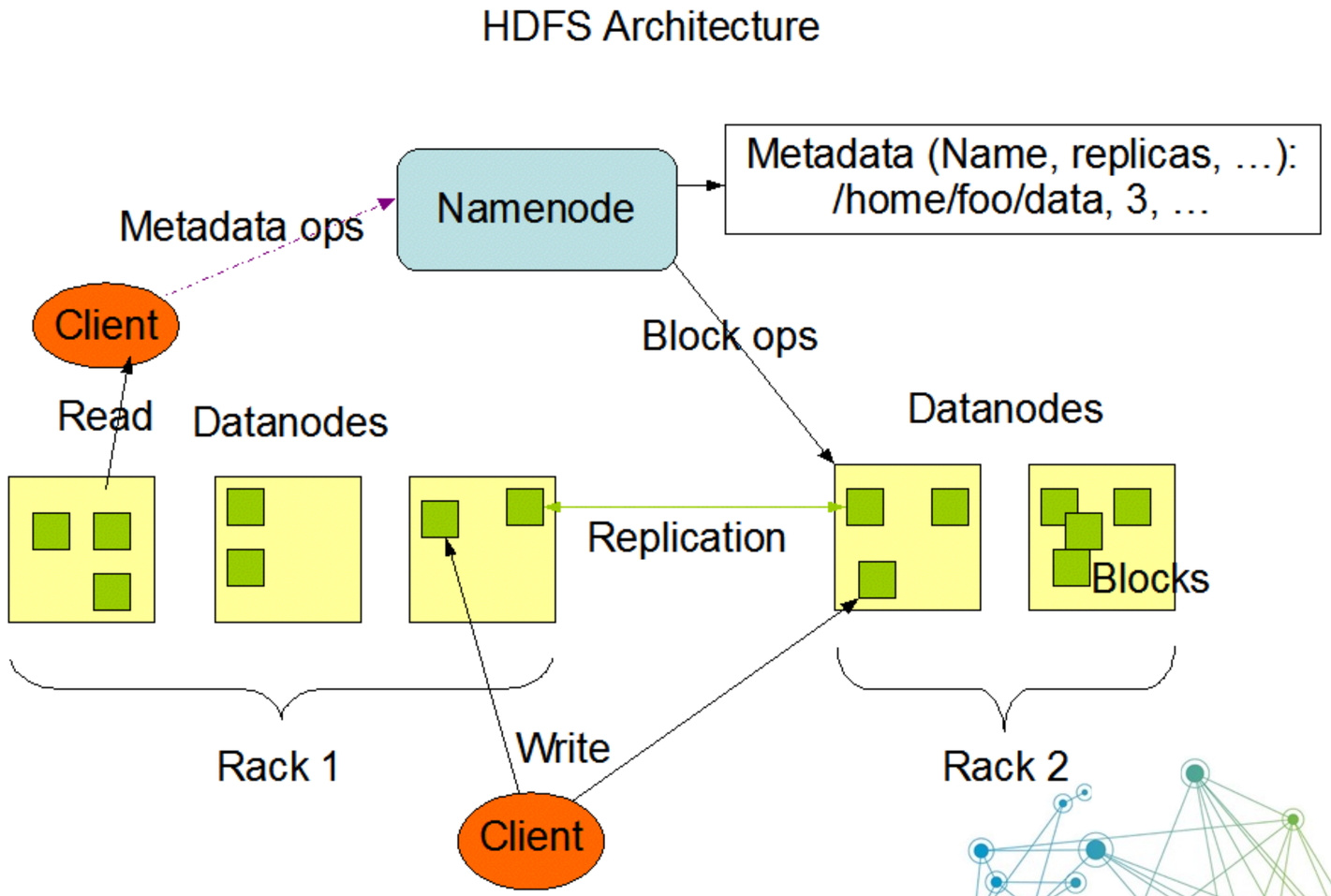
Bulk Data Tools

Bulk Data Framework

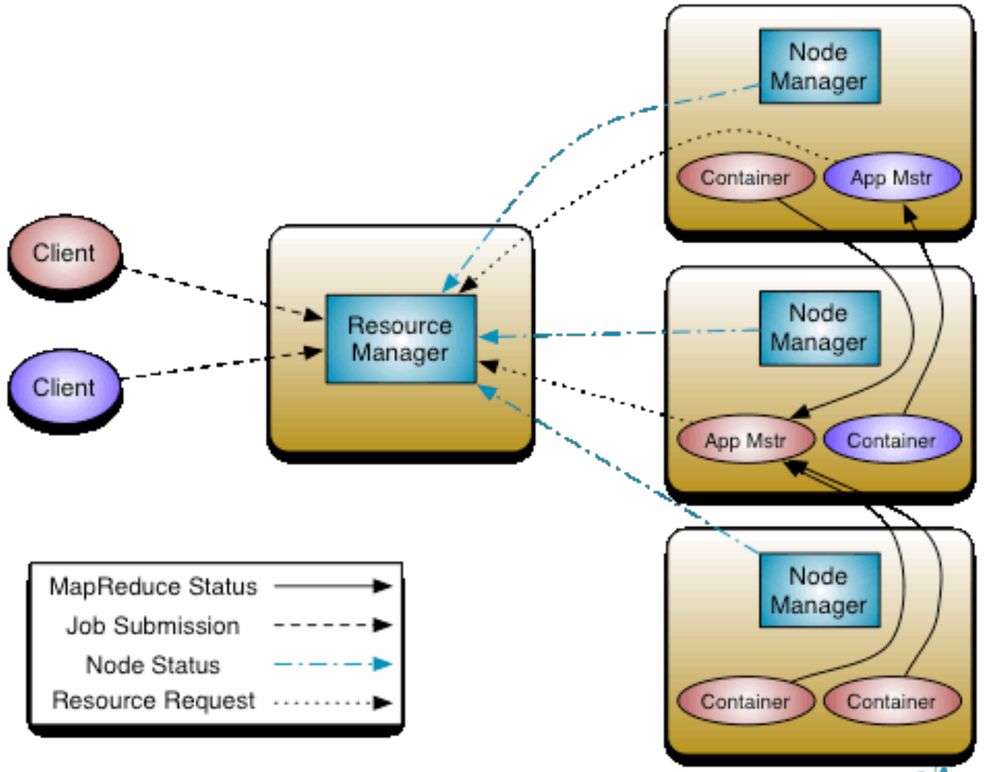
Streaming Framework

Resource Management
(optional)

Bulk / Offline Analysis: HDFS

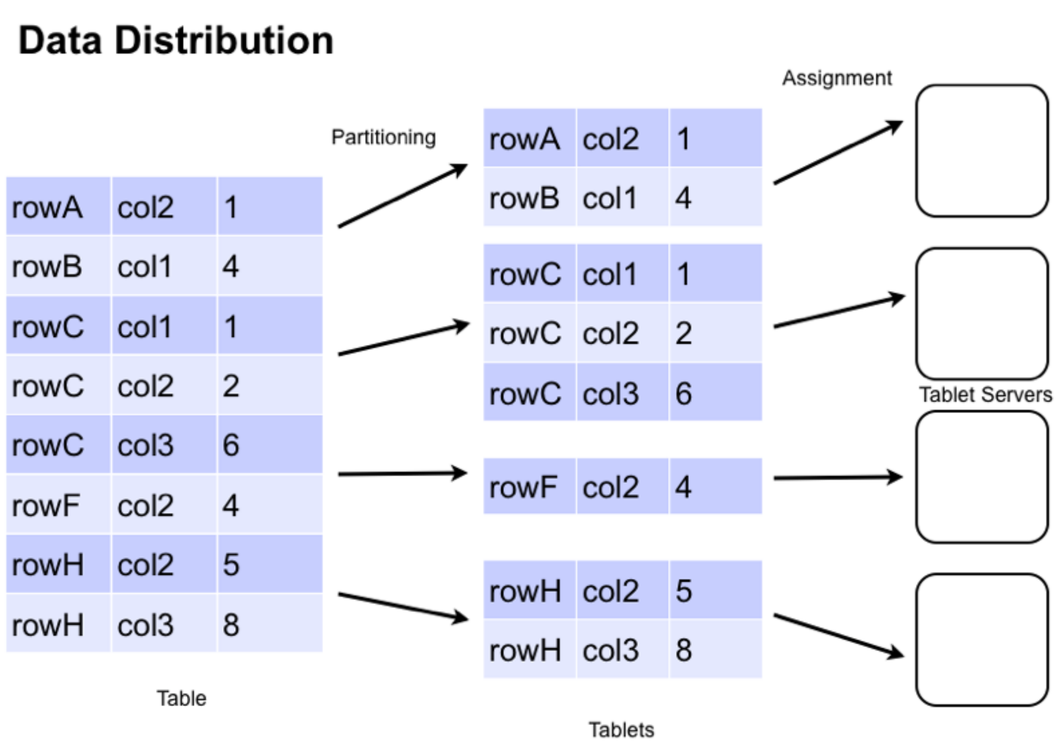


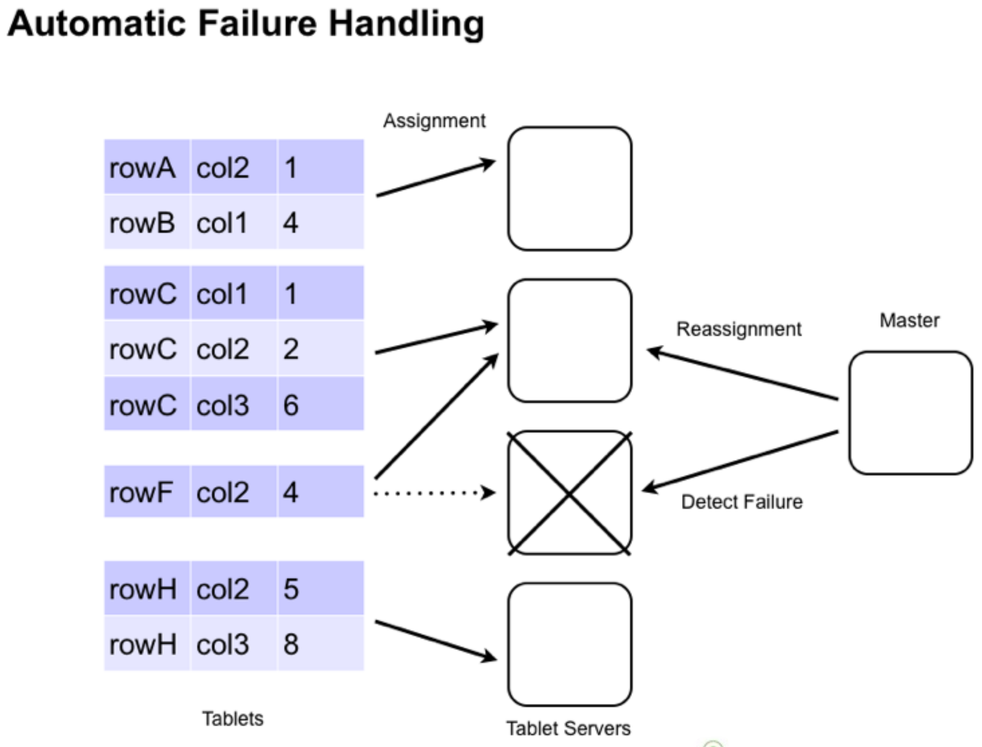
Bulk / Offline Analysis: YARN



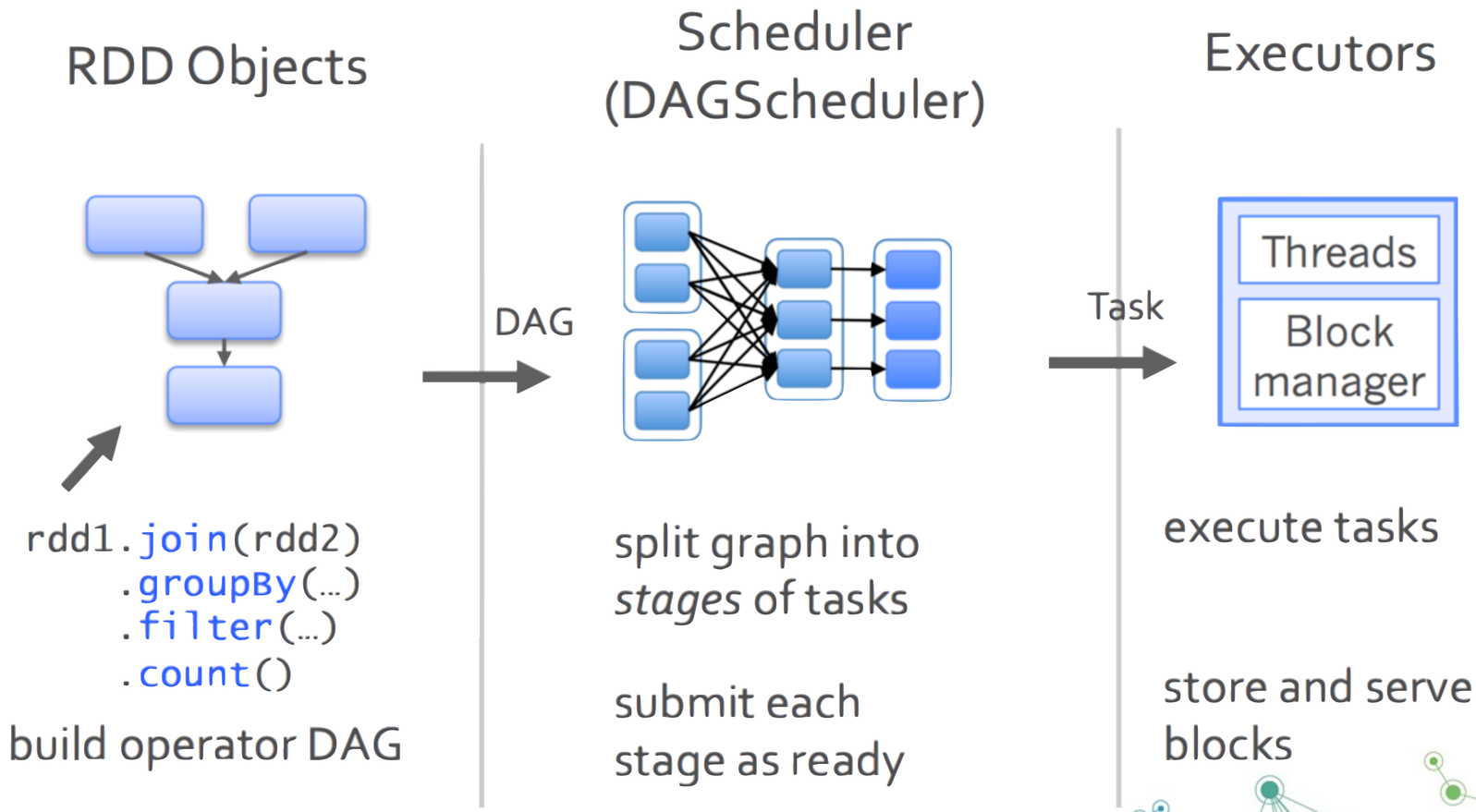
THE UNIVERSITY OF WARWICK

Bulk / Offline Analysis: HBase / Accumulo

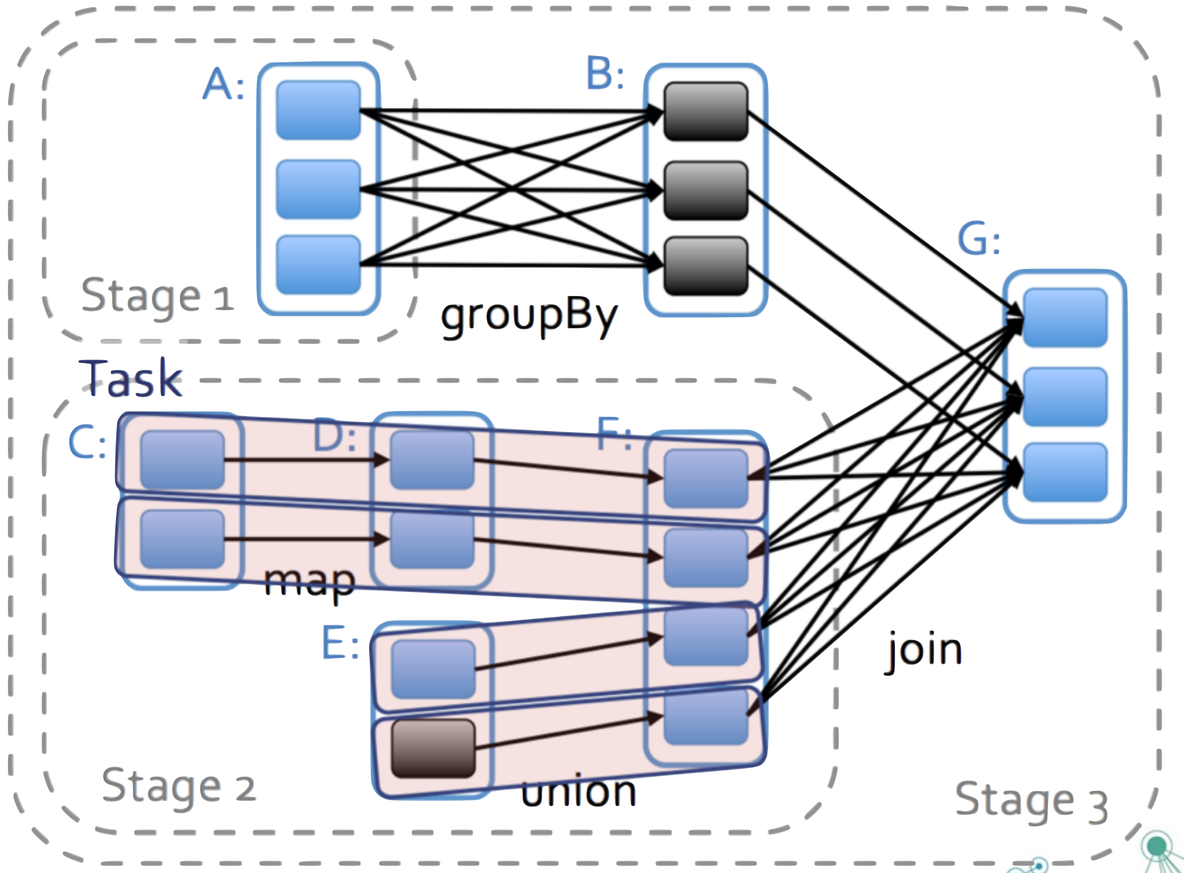




Bulk / Offline Analysis: Spark



Bulk / Offline Analysis: Spark



Bulk / Offline Analysis

Ingest

kafka

Persist

hadoop

APACHE HBASE

accumulo™

Analyse

Spark

Flink

Streaming / Online Analysis

Ingest

kafka

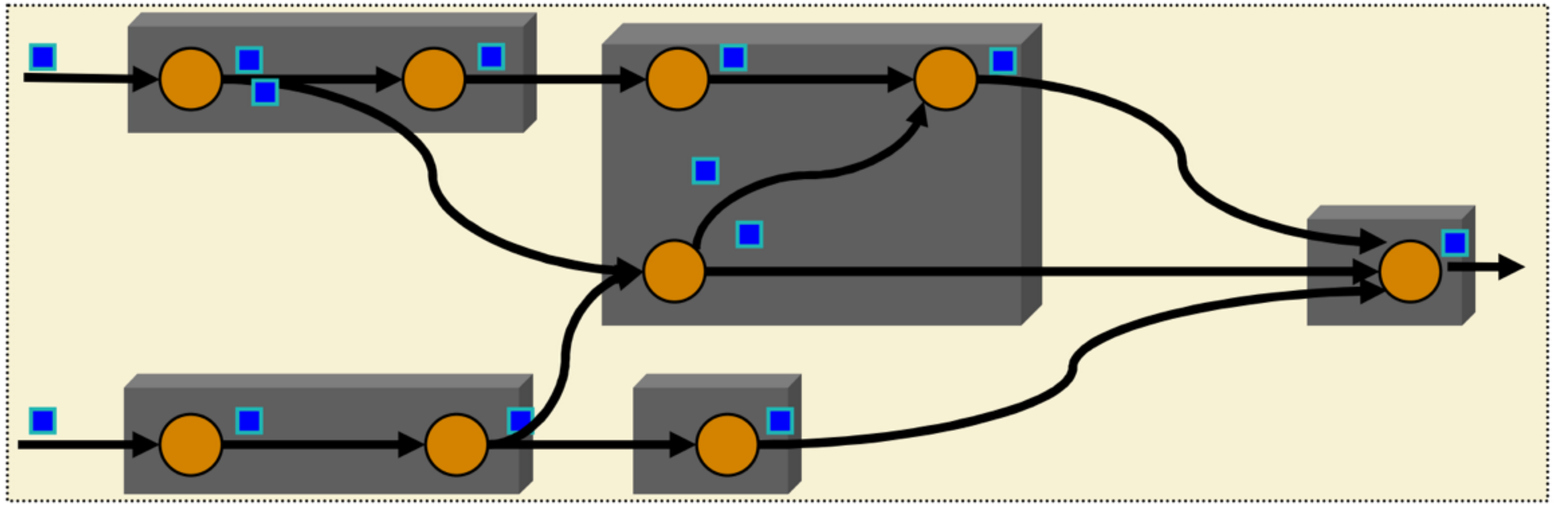
STORM

Spark

InfoSphere Streams

Streaming / Online Analysis

Graph of Processing Elements (PEs):



- Filter
- Aggregate
- Sort
- Join
- Load Shedding
- Custom Functors
- Tool Libraries
- Import / Export (TCP, HDFS, Flat Files, WebSocket, …)

THE UNIVERSITY OF WARWICK

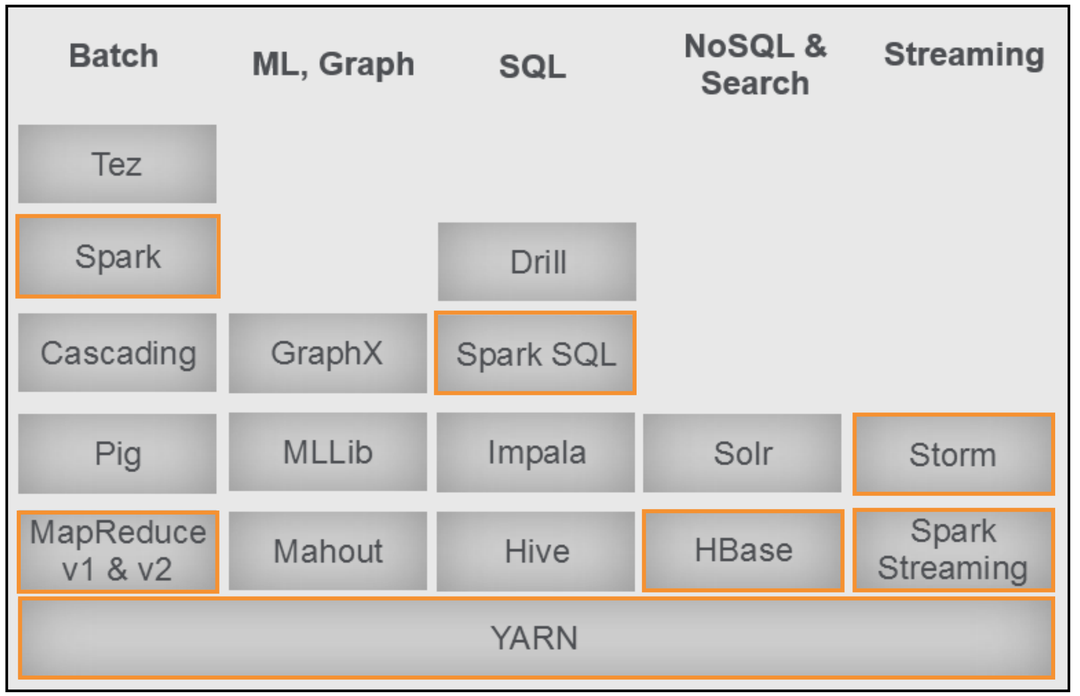
Chiron: WISC/MathSys CDT analytics system

THE UNIVERSITY OF WARWICK

Chiron: WISC/MathSys CDT analytics system

Open architecture multi-platform analytical capability

- Bulk: MapR Hadoop
  - 28x Nodes
  - Mix of HDDs (6TB) and SSDs (3TB)
  - 156TB total storage
- Streaming: Apache Storm
  - 32x Nodes: 64 GB RAM, 20-cores
  - Total 2TB RAM, 640 CPU cores

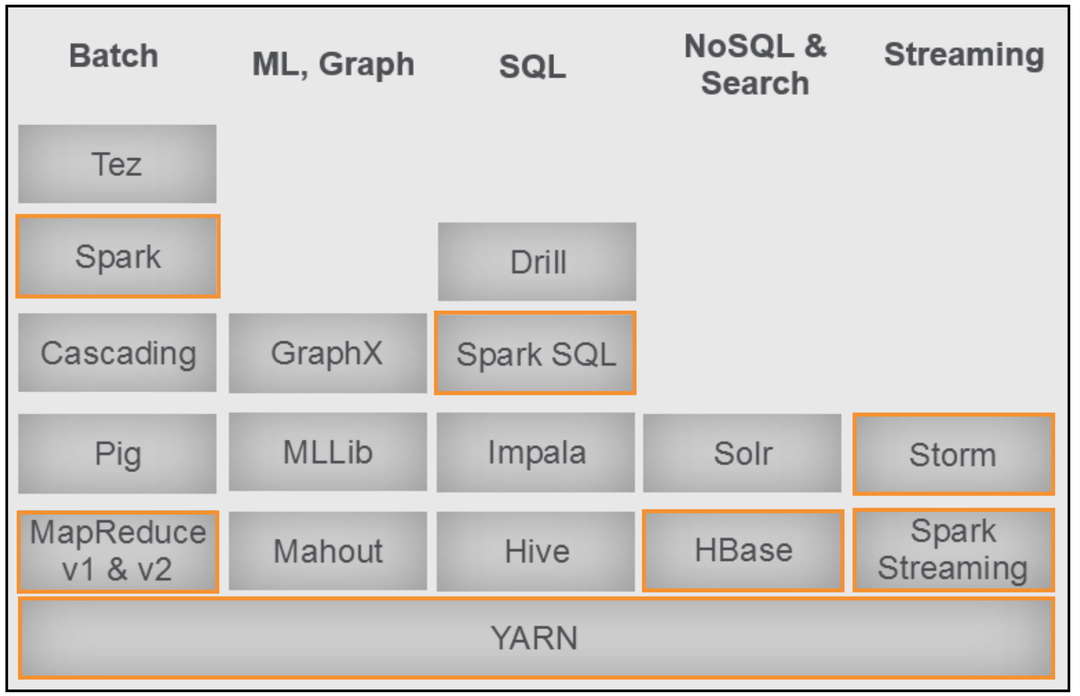


THE UNIVERSITY OF WARWICK

Chiron: WISC/MathSys CDT analytics system

Open architecture multi-platform analytical capability

- Accelerator research platform
  - 2x NVidia K40 Nodes (12GB)
  - 2x Intel Xeon Phi Nodes (16GB)
  - 2x Nallatech 395 FPGA Nodes (32GB)
- Multi-TB in-memory analytics
  - 1x 4TB RAM, 48-core system
- Managed with SLURM



THE UNIVERSITY OF WARWICK

Chiron: WISC/MathSys CDT analytics system

- Shared InfiniBand interconnect: utilise the collection of capabilities as a cohesive whole
- Heterogeneous data-intensive compute applied to the most complex problems
- For example, network activity profiling:
  - Front-end data selection on FPGA
  - Apply a model for traffic classification in parallel on streaming nodes
  - Persistence in HDFS
  - Offline model building in high-memory system

Example analytics applications: Hadoop

Variety of bulk data analytics:

- Telecommunications
  - Motion of citizens around city
  - Disease spread
- Data-driven real-estate valuation
  - Volume & Variety of data
- Parking analysis for urban planning
  - 20,000 terminals in NYC
  - Pattern of life analysis
  - Variable pricing
  - Capacity planning

Example analytics applications: Streaming

Partnership on V2X research programme – streaming onboard sensor data from “Vehicle To X”

- Increase safety and efficiency through enhanced autonomy and cooperation

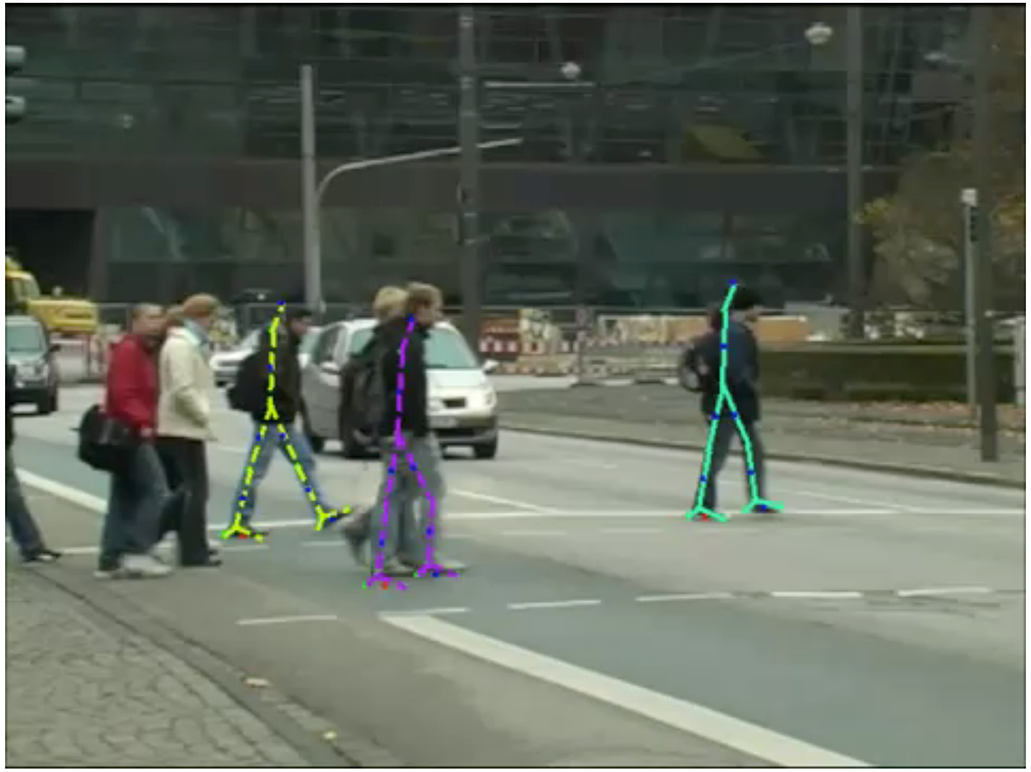
- Streaming sensor data analysis for:
  - City-level traffic management
  - Cooperative navigation
  - Road quality – “Pothole Patrol”
  - Predictive analytics for vehicle maintenance

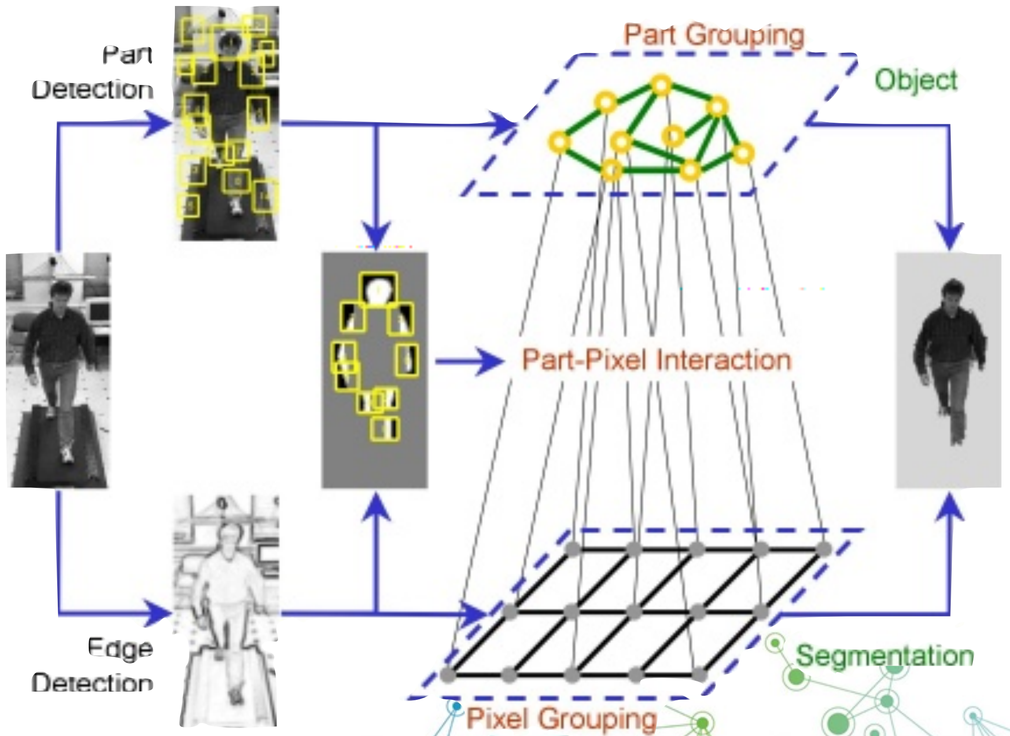
Example analytics applications: Accelerators

- Proven track record in HPC simulation – at the intersection of HPC and Big Data
- Early days for data intensive applications on accelerators
- Opportunity to apply GPGPU; Xeon Phi; FPGA to data intensive applications including:
  - Multi-scale economic simulations
  - Offloading compute intensive analysis
  - Cybersecurity
  - Machine learning

Example analytics applications: Accelerators

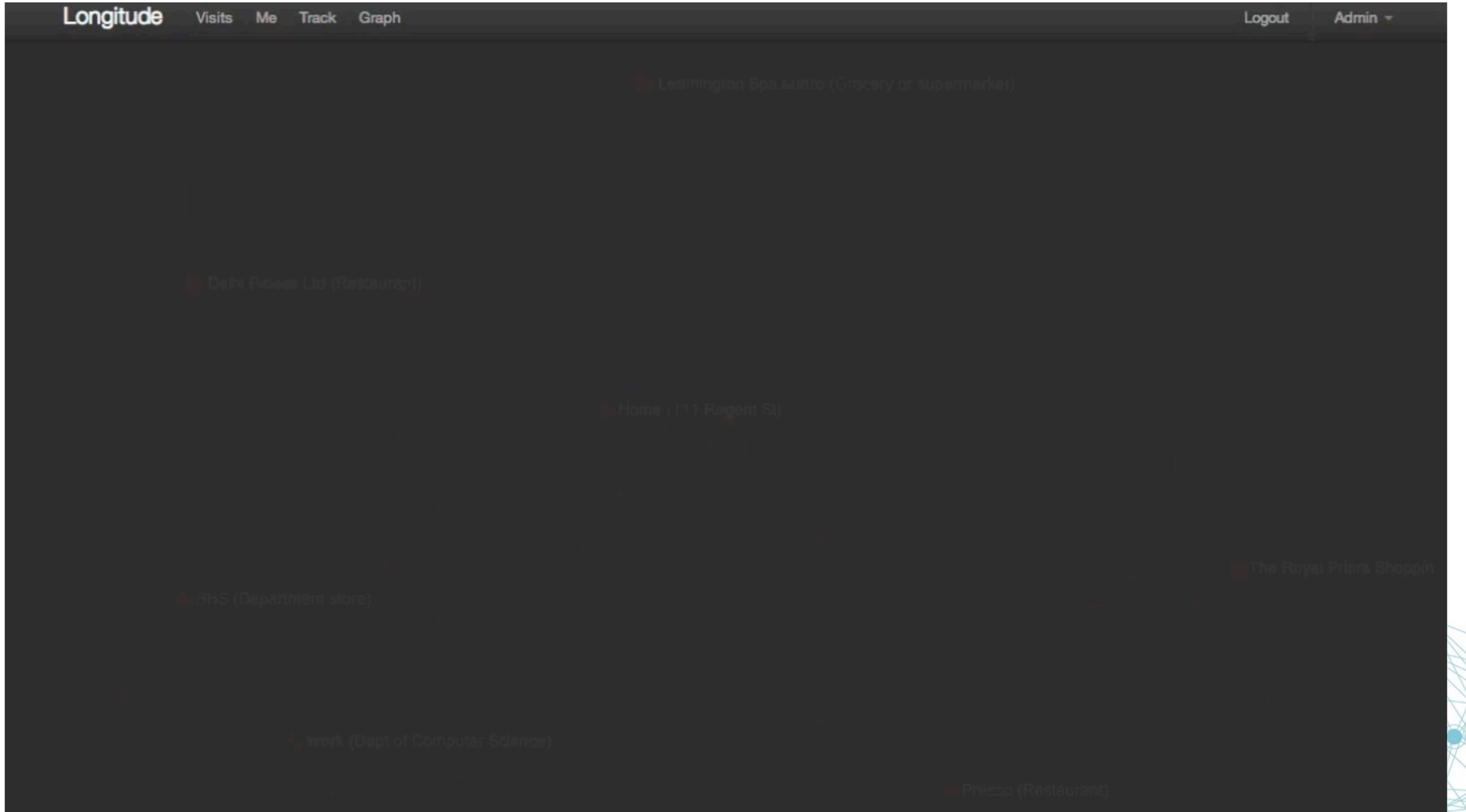
Near-real-time gait analysis, recognition, and tracking





Example analytics applications: High Memory

Large-scale graph analytics and machine learning problems: Pattern of life in an urban context



THE UNIVERSITY OF WARWICK

Chiron: WISC/MathSys CDT analytics system

Heterogeneous data-intensive compute, applied to the most complex real-world problems


QUESTIONS?
http://hadoop.apache.org/
http://accumulo.apache.org/
http://spark.apache.org/
http://storm.apache.org/
http://slurm.schedmd.com/
https://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-developers-quick-start-guide
http://docs.nvidia.com/cuda/https://www.altera.com/products/design-software/overview.html

WARWICK INSTITUTE
FOR THE
SCIENCE OF CITIES

www.wisc.warwick.ac.uk

THE UNIVERSITY OF
WARWICK

Warwick Institute for the Science of Cities
The University of Warwick, CV4 7AL