BIG DATA EUROPE

Empowering Communities
with Data Technologies

# Collaboration Opportunities in Big Data Platform, Data Spaces & Semantic Interoperability

Sören Auer, auer@cs.uni-bonn.de

PICASSO EU-US Collaboration, Minneapolis, June 19, 2017

# PICASSO ICT MEETING

Three possible streams for collaboration:

- A Big Data Platform for societal good
- Establishing data sharing and data value chains with the Industrial Data Space
- Semantic Domain Models (vocabularies, ontologies) for establishing a common understanding of the data

# Big Data Europe Platform

Empowering Communities with Data Technologies
Platform release

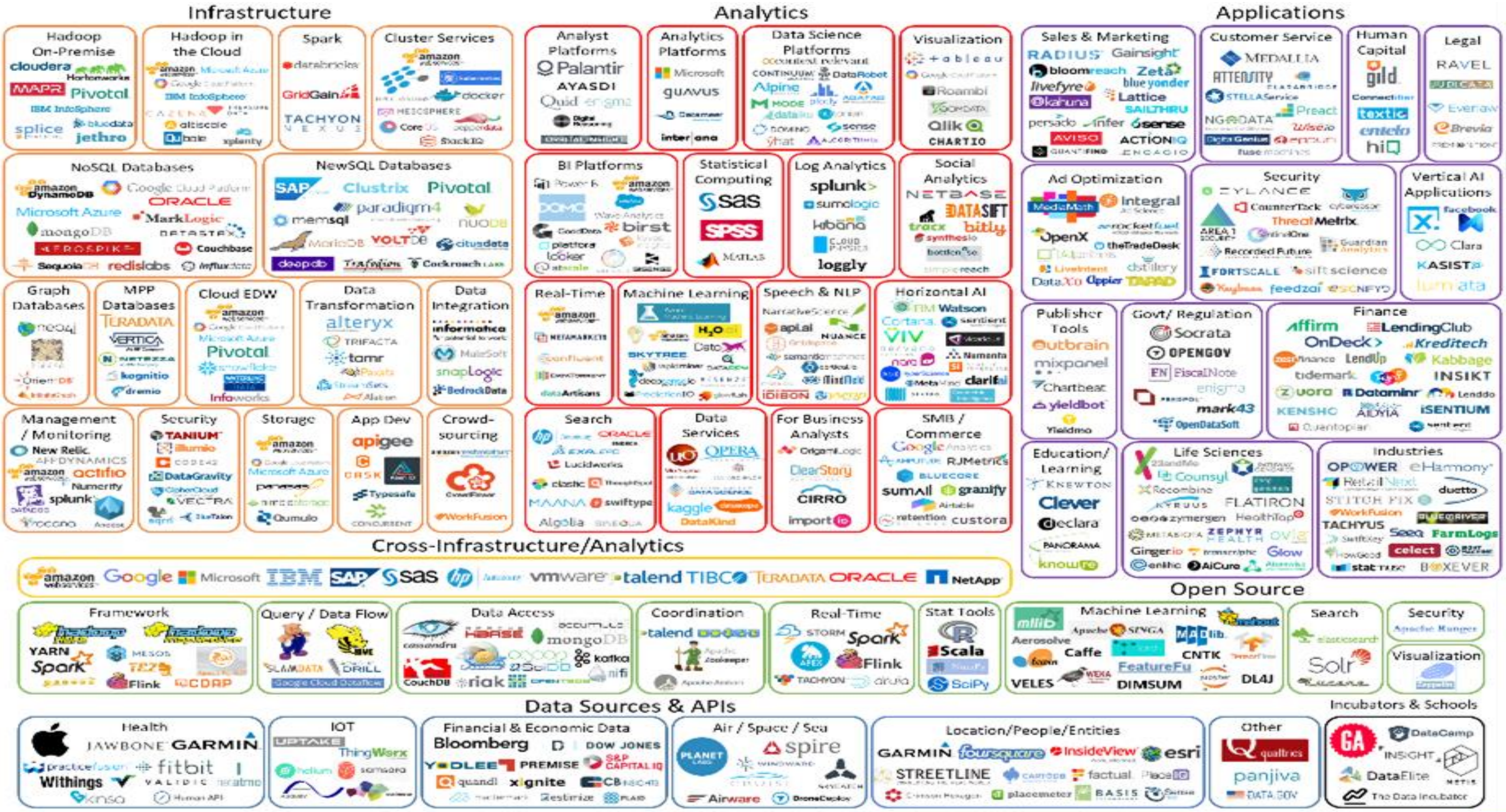
BIG DATA EUROPE

Empowering Communities
with Data Technologies

Smart Data Analytics

# Platform Goals

- Opensource
- Ease of Use
- Support a variety of use cases
- Embrace emerging Big Data Technologies
- Simple integration with custom components



## Infrastructure

Hadoop On-Premise · Hadoop in the Cloud · Spark · Cluster Services · NoSQL Databases · NewSQL Databases · Graph Databases · MPP Databases · Cloud EDW · Data Transformation · Data Integration · Management / Monitoring · Security · Storage · App Dev · Crowd-sourcing

## Analytics

Analyst Platforms · Analytics Platforms · Data Science Platforms · Visualization · BI Platforms · Statistical Computing · Log Analytics · Social Analytics · Real-Time · Machine Learning · Speech & NLP · Horizontal AI · Search · Data Services · For Business Analysts · SMB / Commerce

## Applications

Sales & Marketing · Customer Service · Human Capital · Legal · Ad Optimization · Security · Vertical AI Applications · Publisher Tools · Govt/ Regulation · Finance · Education/ Learning · Life Sciences · Industries

## Cross-Infrastructure/Analytics

## Open Source

Framework · Query / Data Flow · Data Access · Coordination · Real-Time · Stat Tools · Machine Learning · Search · Security · Visualization

## Data Sources & APIs

Health · IOT · Financial & Economic Data · Air / Space / Sea · Location/People/Entities · Other

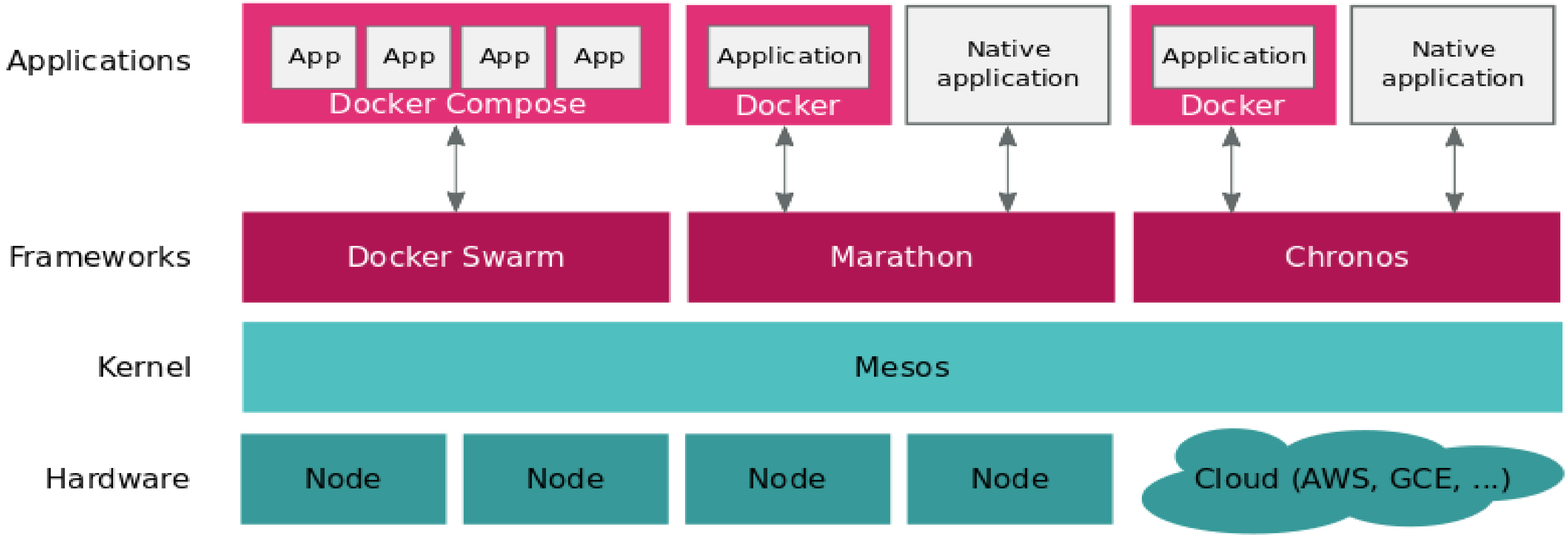
Incubators & Schools

cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, splice, bluedata, jethro
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Qubole, xplenty
databricks, GridGain, TACHYON NEXUS
amazon web services, docker, MESOSPHERE, CoreOS, StackIQ
Palantir, AYASDI, Quid, enigma
Microsoft, guavus, Datameer, interana
continuum, DataRobot, Alpine, MODE, plotly, dataiku, DOMINO, sense, yhat
tableau, Google Cloud Platform, Roambi, Qlik, CHARTIO
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, redislabs, influxdata
SAP, Clustrix, Pivotal, paradigm4, memsql, NuoDB, MariaDB, VoltDB, citusdata, deepdb, Trafodion, CockroachLabs
Power BI, amazon, DOMO, Wave Analytics, GoodData, birst, platfora, looker, atscale
sas, SPSS, MATLAB
splunk, sumologic, kibana, loggly
NETBASE, DATASIFT, bitly, synthesio, simplereach
neo4j, OrientDB
TERADATA, VERTICA, NETEZZA, kognitio, dremio
amazon, Microsoft Azure, Pivotal, snowflake, Infoworks
alteryx, TRIFACTA, tamr, Paxata
informatica, MuleSoft, snapLogic, BedrockData
amazon, METAMARKETS, confluent, dataArtisans
Azure Machine Learning, H2O, SKYTREE, DataRobot
NarrativeScience, api.ai, NUANCE, idibon
IBM Watson, Cortana, sentient, VIV, Numenta, clarifai
New Relic, APPDYNAMICS, amazon, actifio, Numerify, splunk, DATADOG
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl
amazon, Microsoft Azure, panasas, Qumulo
apigee, CASK, Typesafe, CONCURRENT
CrowdFlower, WorkFusion
hp, ORACLE, EXALEAD, Lucidworks, elastic, MAANA, swiftype, Algolia
OPERA, kaggle, DataKind
ClearStory, CIRRO, import.io
Google Analytics, RJMetrics, BLUECORE, sumAll, granify, custora
RADIUS, Gainsight, bloomreach, Zeta, livefyre, blue yonder, kahuna, Lattice, SAILTHRU, persado, infer, 6sense, AVISO, ACTION IQ
MEDALLIA, ATTENSITY, STELLAService, NGDATA, Preact, Wise.io
gild, Connectifier, textio, entelo, hiQ
RAVEL, JUDICATA, Everlaw, Brevia
MediaMath, Integral, OpenX, rocketfuel, theTradeDesk, DataXu, Applier, TAPAD
CYLANCE, CounterTack, Cybereason, AREA 1, ThreatMetrix, Recorded Future, FORTSCALE, sift science, feedzai
X.ai, facebook, Clara, KASISTO
Outbrain, mixpanel, Chartbeat, yieldbot
Socrata, OPENGOV, FiscalNote, enigma, mark43, OpenDataSoft
Affirm, LendingClub, OnDeck, Kreditech, Kabbage, LendUp, tidemark, INSIKT, zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM
KNEWTON, Clever, declara, PANORAMA, knowre
23andMe, Counsyl, Recombine, FLATIRON, zymergen, HealthTap, ZEPHYR HEALTH, Ginger.io, Enlitic, AiCure
OPOWER, eHarmony, Retail Next, duetto, STITCH FIX, WorkFusion, TACHYUS, BLUE RIVER, FarmLogs, select, stat, BOXEVER
amazon web services, Google, Microsoft, IBM, SAP, sas, hp, vmware, talend, TIBCO, TERADATA, ORACLE, NetApp
hadoop, YARN, Spark, MESOS, TEZ, Flink, CDAP
SLAMDATA, DRILL, Google Cloud Dataflow
cassandra, HBASE, mongoDB, accumulo, CouchDB, riak, kafka, nifi
talend, Apache Zookeeper, Apache Ambari
STORM, Spark, APEX, Flink, TACHYON, druid
R, Scala, SciPy
mllib, Apache SINGA, MADlib, mahout, Aerosolve, Caffe, CNTK, FeatureFu, VELES, DIMSUM, DL4J
elasticsearch, Solr, Lucene
Apache Ranger
JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API
UPTAKE, ThingWorx, helium, samsara
Bloomberg, DOW JONES, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB Insights, estimize
PLANET LABS, spire, WINDWARD, Airware, DroneDeploy
GARMIN, foursquare, InsideView, esri, STREETLINE, factual, Place IQ, placemeter, BASIS
qualtrics, panjiva, DATA.GOV
GA, DataCamp, INSIGHT, METIS, DataElite, The Data Incubator

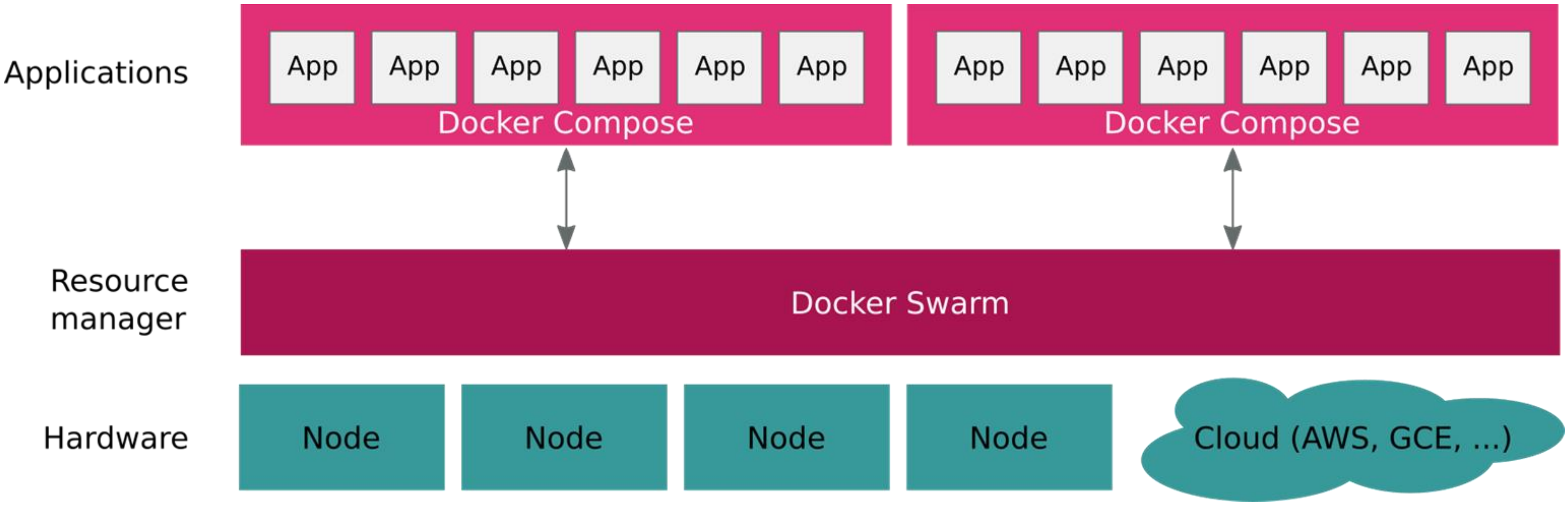
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

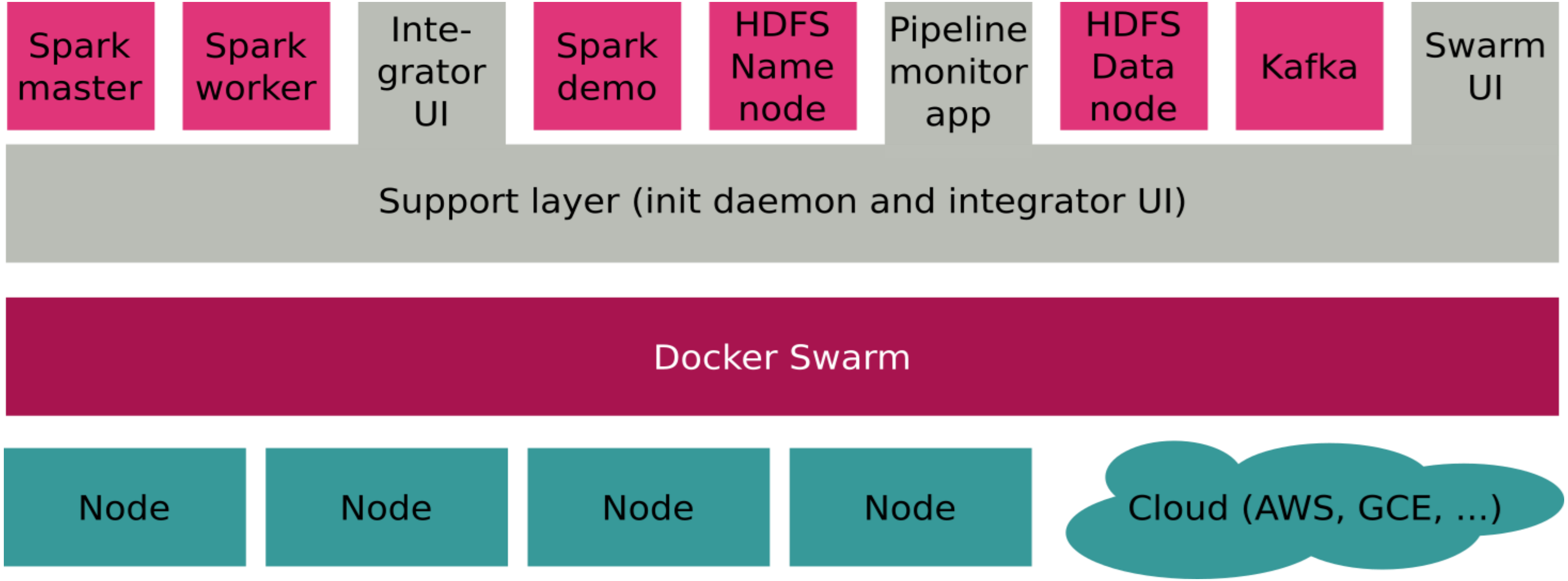
# Platform Architecture Evolution



# Platform Architecture Evolution



# Platform Architecture Existing



Spark master

Spark worker

Integrator UI

Spark demo

HDFS Name node

Pipeline monitor app

HDFS Data node

Kafka

Swarm UI

Support layer (init daemon and integrator UI)

Docker Swarm

Node

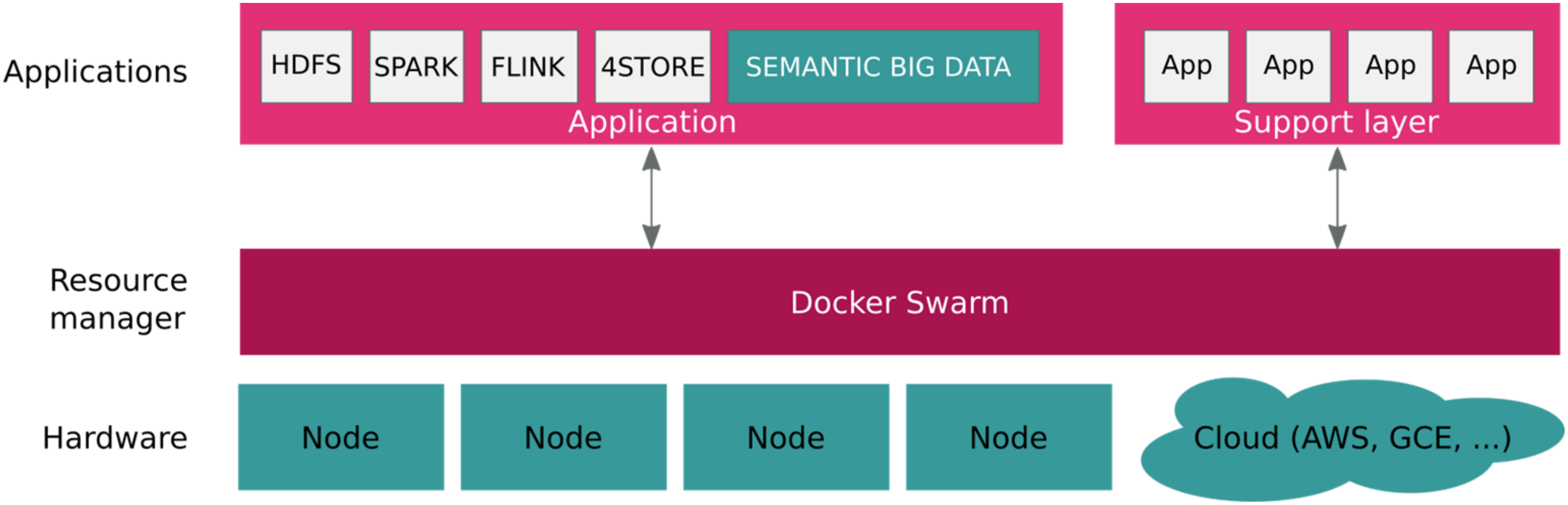
Node

Node

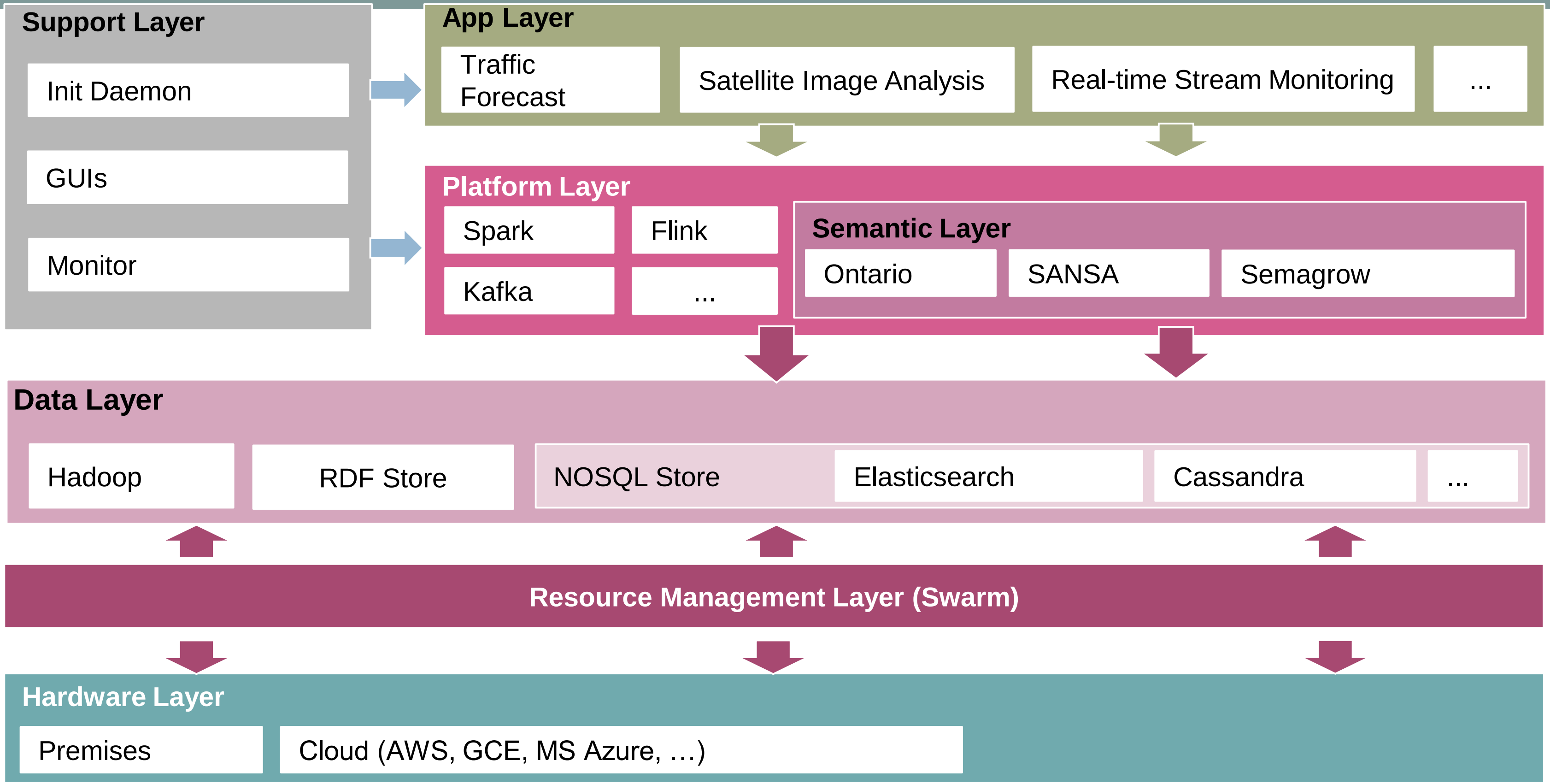
Node

Cloud (AWS, GCE, ...)

# Platform Architecture Existing



# Platform Architecture Alternate View



Support Layer

Init Daemon

GUIs

Monitor

App Layer

Traffic Forecast

Satellite Image Analysis

Real-time Stream Monitoring

...

Platform Layer

Spark

Flink

Kafka

...

Semantic Layer

Ontario

SANSA

Semagrow

Data Layer

Hadoop

RDF Store

NOSQL Store

Elasticsearch

Cassandra

...

Resource Management Layer (Swarm)

Hardware Layer

Premises

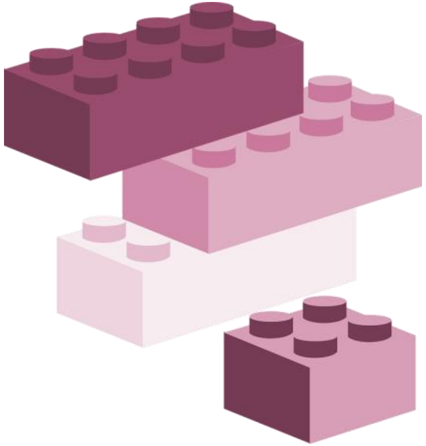
Cloud (AWS, GCE, MS Azure, …)

# BDE Supported Frameworks

| **Search/indexing** | **Data processing** |
|---|---|
| Apache Solr | Apache Spark |
| **Data acquisition** | Apache Flink |
| Apache Flume | **Semantic Components** |
| **Message passing** | Strabon |
| Apache Kafka | Sextant |
| **Data storage** | GeoTriples |
| Hue | Silk |
| Apache Cassandra | SEMAGROW |
| ScyllaDB | LIMES |
| Apache Hive | 4Store |
| Postgis | OpenLink Virtuoso |

# Platform features

- BDE Development Environment
  - Stack builder
  - Workflow builder
  - Instructions to add custom components to the BDE stack
- Administrator Interface
  - SwarmUI
- UI Integrator
  - Workflow monitor
  - Integrated web interface



# What BDE Provides ?

- Platform Installation Instructions
- Usage Instructions
  - Creating a stack
  - Creating a workflow
  - Monitoring the Stack
  - Integration of Custom Components

# Platform installation

- Manual installation guide
- Using Docker Machine
  - On local machine (VirtualBox)
  - In cloud (AWS, DigitalOcean, Azure)
  - Bare metal
- Screencasts

https://www.big-data-europe.eu/platform/

https://github.com/big-data-europe

# Deploying a Big Data Stack

- Stack Builder
- Stack
  - Collection of communicating components to solve a specific problem
- Described in Docker Compose
  - Component configuration
  - Application topology

# Creation of WorkFlows

- Pipeline Builder
  - Allows creation of dependencies among different applications
- WorkFlow Monitor
  - Monitoring of pipeline-workflow using

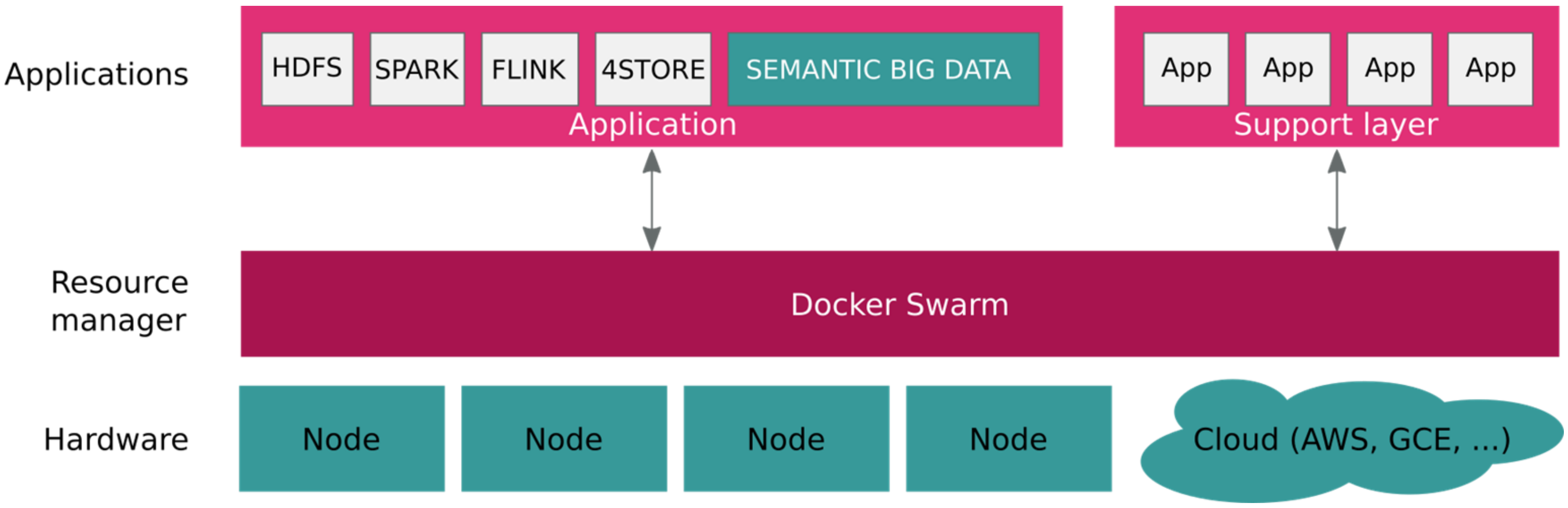
# Integrating Custom Components

- Instructions
  - Orchestrator required for initialization process (init_daemon)
    - Components may depend on each other
    - Components may require manual intervention
  - User Interface Integration
    - Standard Interfaces from components
    - Combine and align the interfaces

# User Interfaces

- Target: Facilitate the use of the platform
  - User Interface Adaption
- Available interfaces
  - Workflow UIs
    - Workflow Builder
    - Workflow Monitor
  - Swarm UI
  - Integrator UI

# Platform Architecture



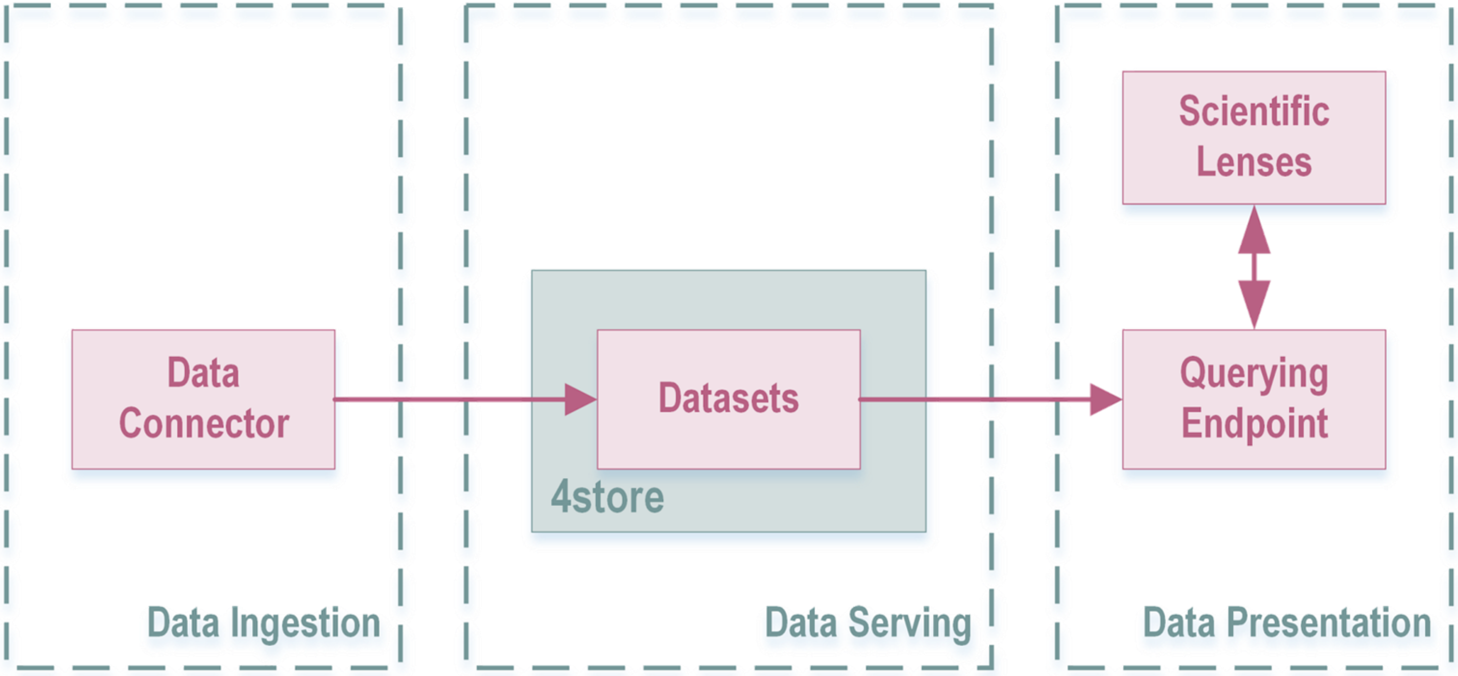
# Pilot Show Cases

SC1 SC2 SC3 SC4 SC5 SC6

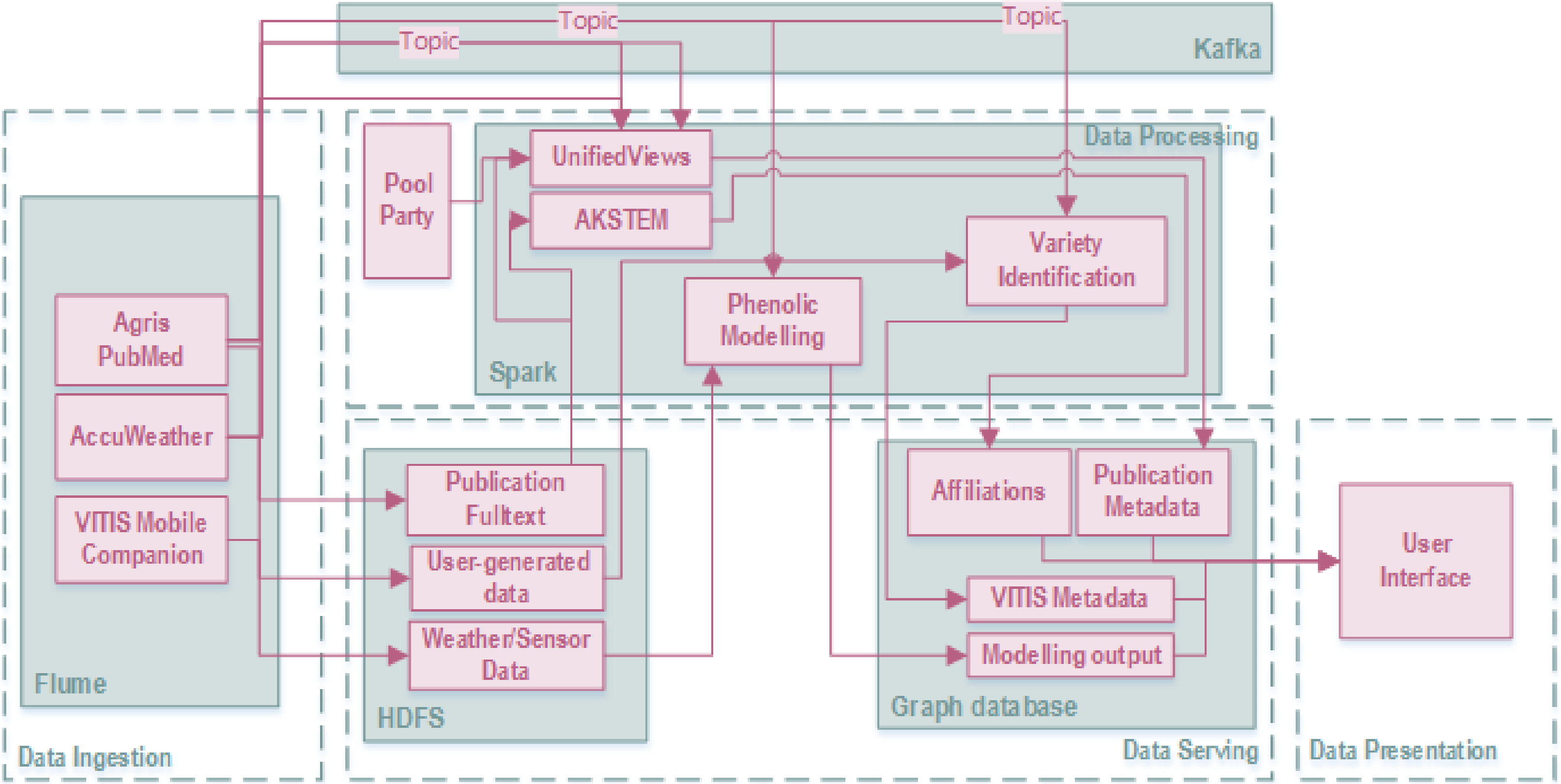
SC7

SC1 - Open PHACTS discovery platform relating to biological/medical questions
SC2 - Discovery and Linking of Viticulture-relevant information
SC3 - System monitoring in energy production units
SC4 - Short-Term traffic flow forecasting.
SC5 - Supporting data-intensive climate research
SC6 - Citizens & Researchers Budget on Municipal Level
SC7 - Ingestion of remote sensing images and social sensing data to detect and verify changes on the Earth surface for security applications

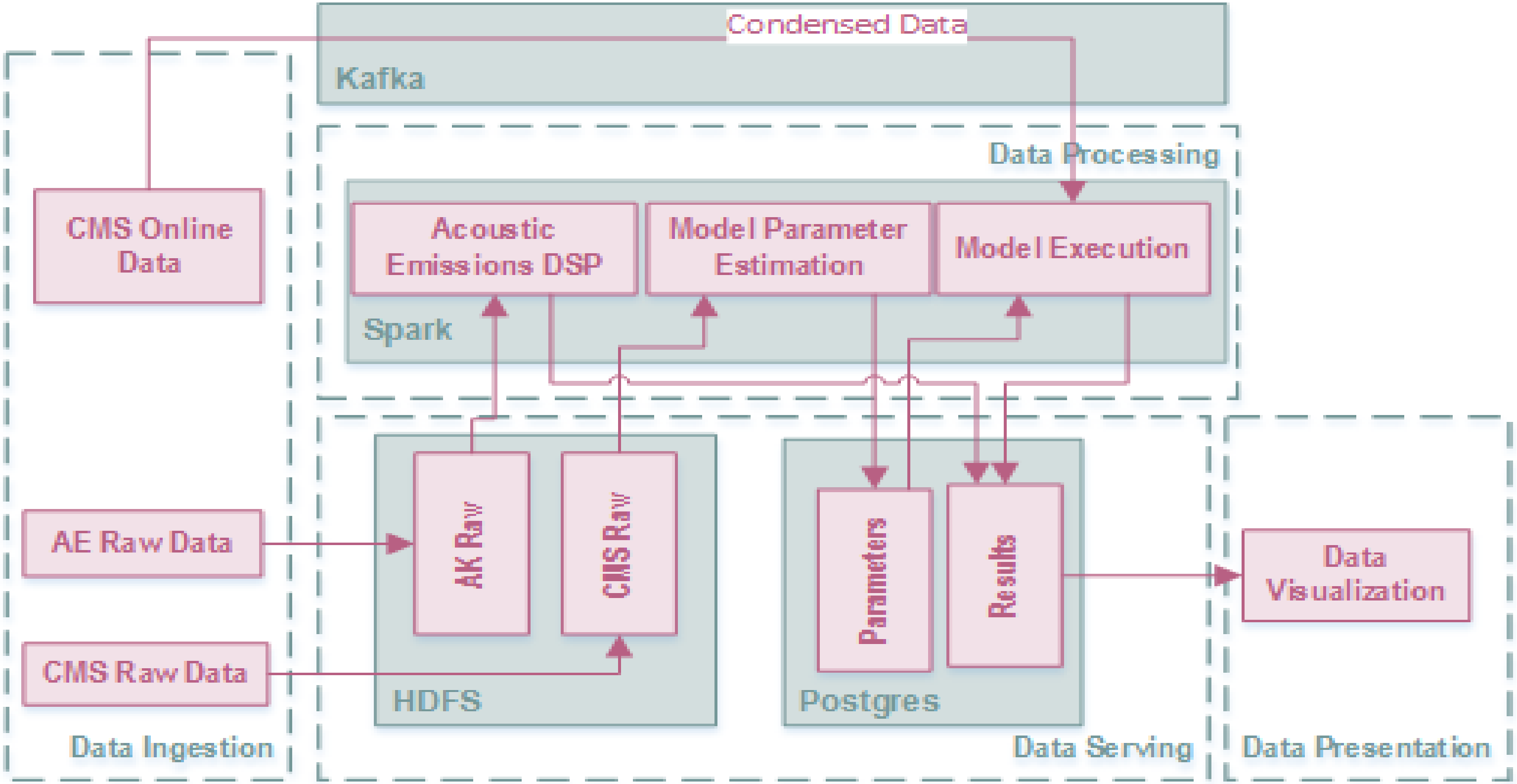
# SC1- Health


Data Connector
Datasets
4store
Scientific Lenses
Querying Endpoint
Data Ingestion
Data Serving
Data Presentation

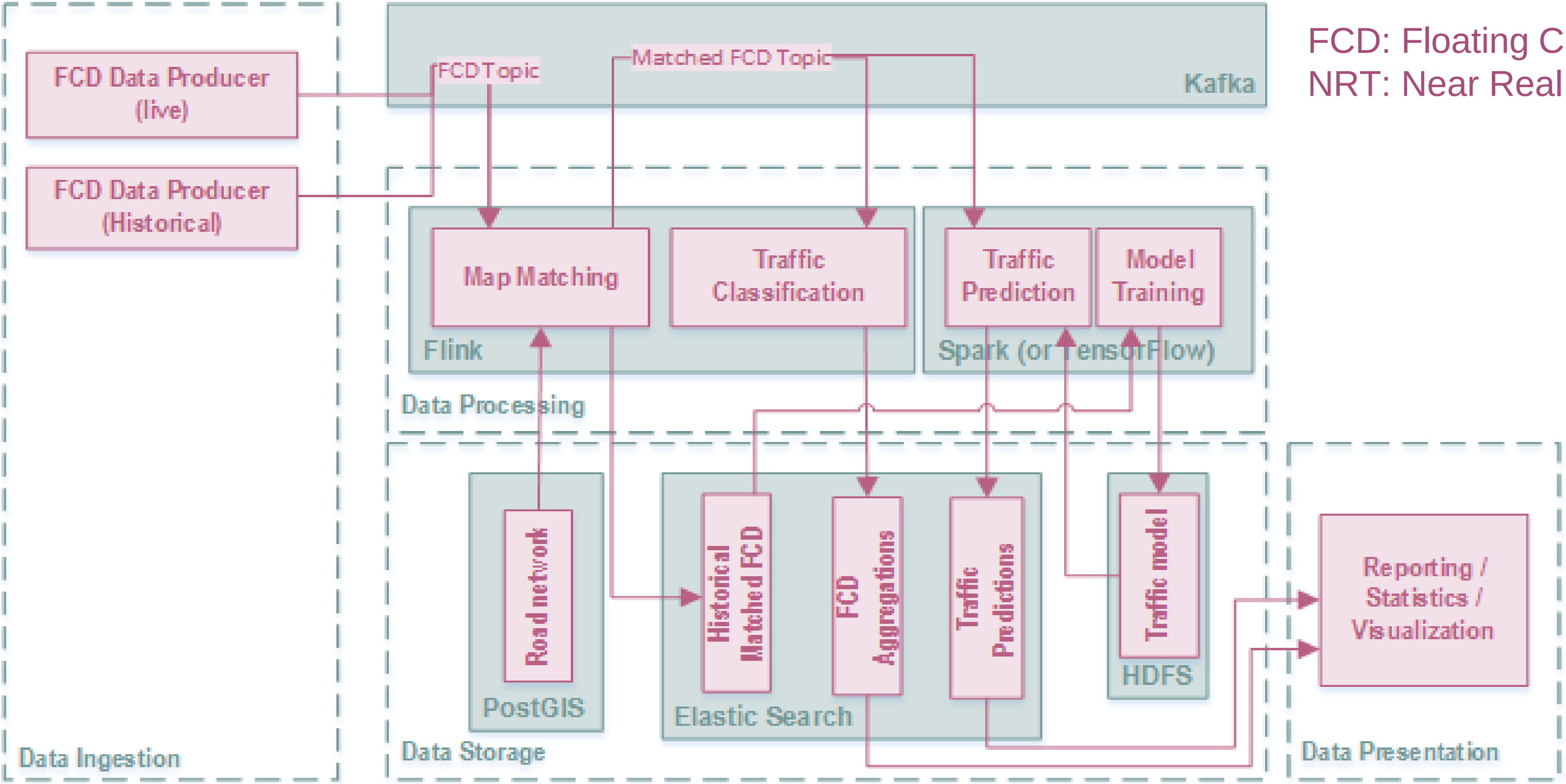
# SC2 - Food


Topic
Topic
Topic
Kafka
Data Processing
Pool Party
UnifiedViews
AKSTEM
Variety Identification
Phenolic Modelling
Spark
Agris PubMed
AccuWeather
VITIS Mobile Companion
Flume
Data Ingestion
Publication Fulltext
User-generated data
Weather/Sensor Data
HDFS
Affiliations
Publication Metadata
VITIS Metadata
Modelling output
Graph database
Data Serving
User Interface
Data Presentation

# SC3 - Energy

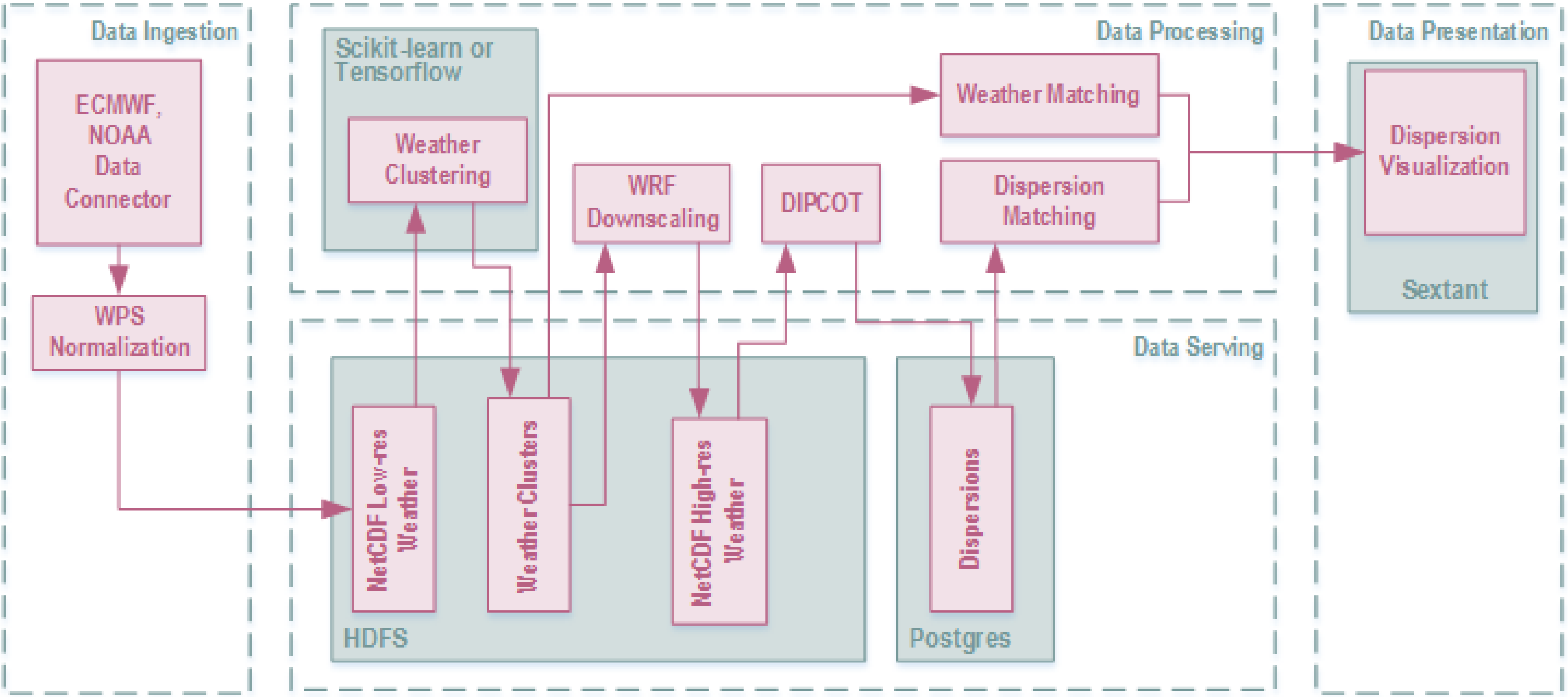

Condensed Data
Kafka
Data Processing
CMS Online Data
Acoustic Emissions DSP
Model Parameter Estimation
Model Execution
Spark
AE Raw Data
CMS Raw Data
Data Ingestion
AK Raw
CMS Raw
HDFS
Parameters
Results
Postgres
Data Serving
Data Visualization
Data Presentation

# SC4 - Transport

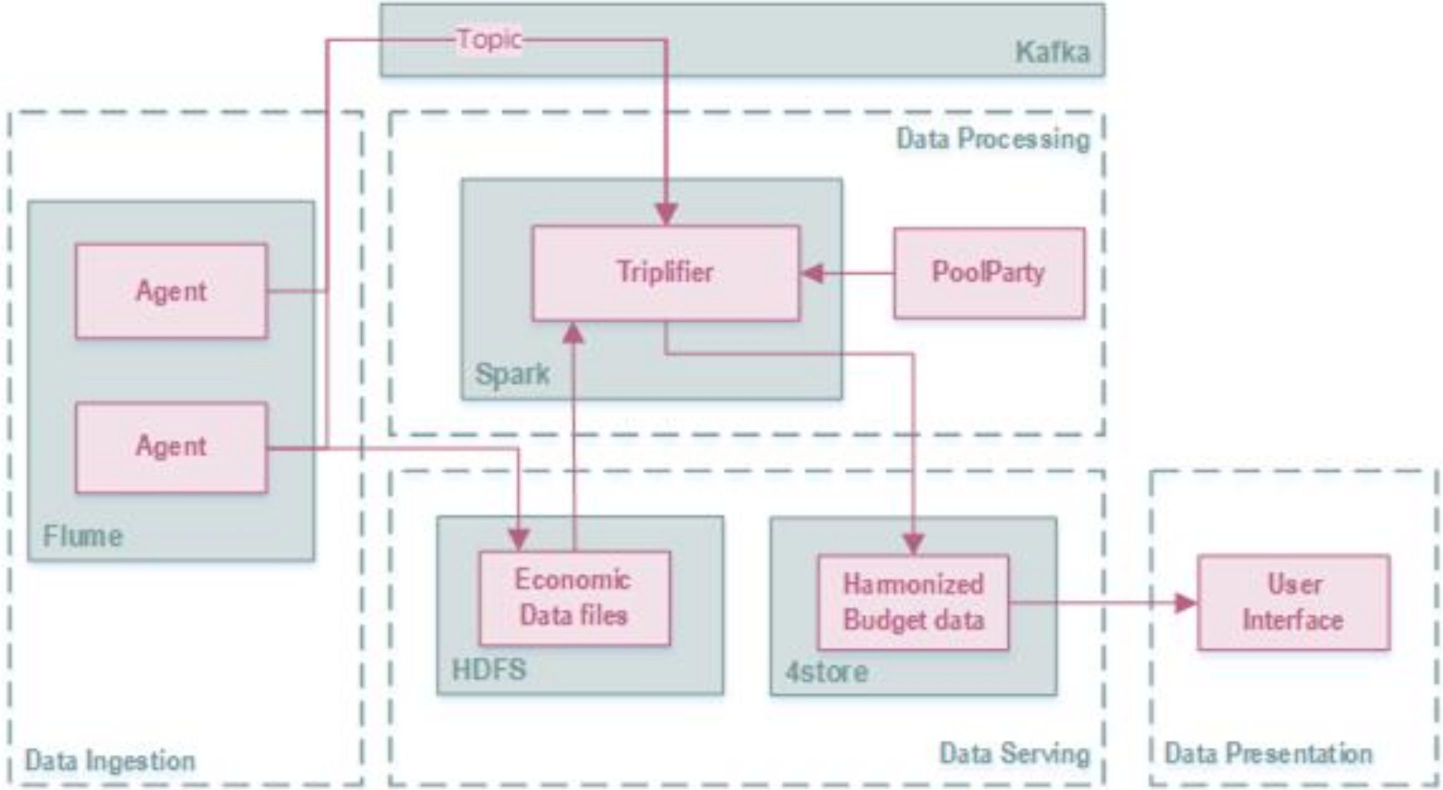


FCD: Floating Car Data
NRT: Near Real Time

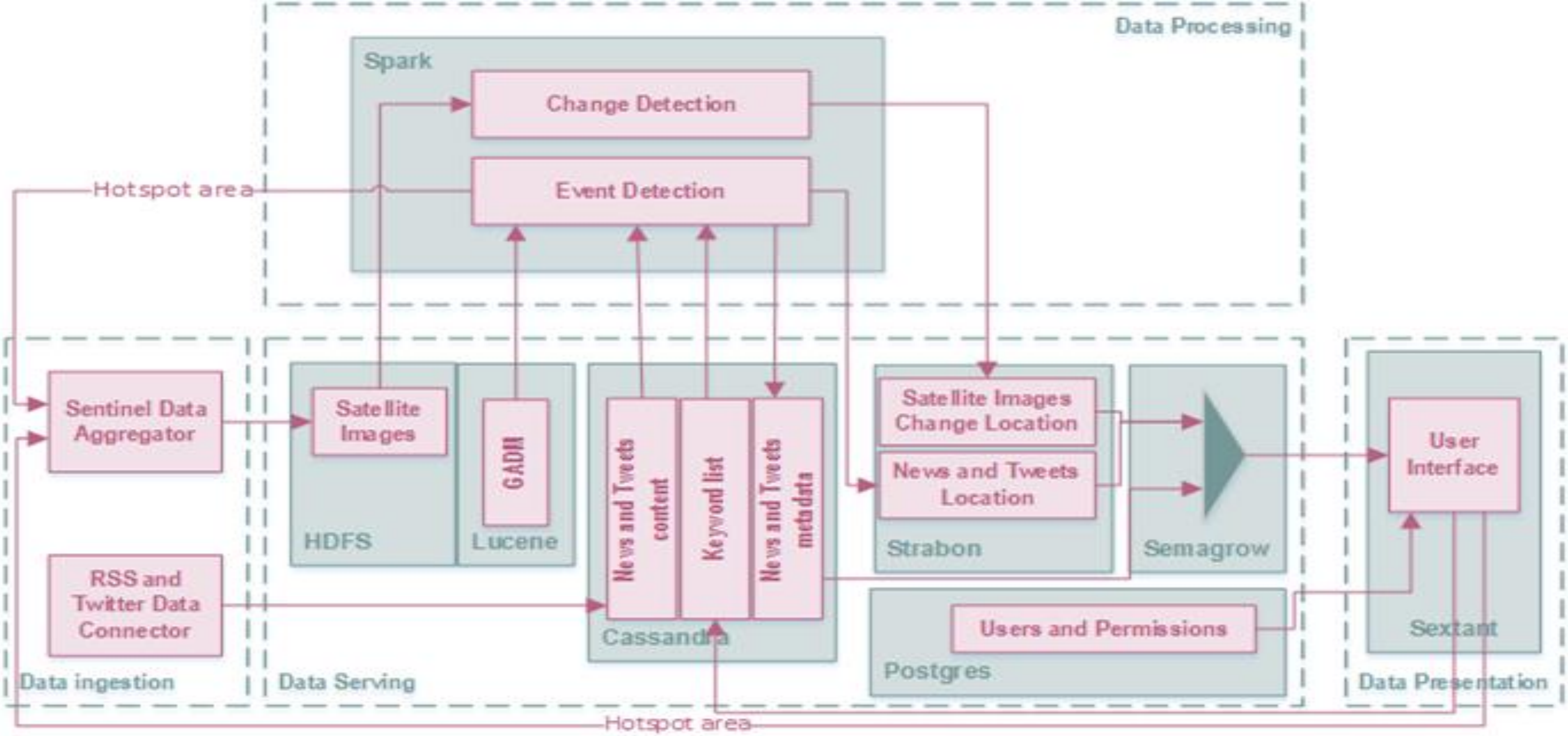
# SC5 - Climate


Data Ingestion
ECMWF, NOAA Data Connector
WPS Normalization
Data Processing
Scikit-learn or Tensorflow
Weather Clustering
WRF Downscaling
DIPCOT
Weather Matching
Dispersion Matching
Data Serving
NetCDF Low-res Weather
Weather Clusters
NetCDF High-res Weather
HDFS
Dispersions
Postgres
Data Presentation
Dispersion Visualization
Sextant

# SC6 - Social Sciences



# SC7 - Security



# BDE vs Hadoop distributions

| | Hortonworks | Cloudera | MapR | Bigtop | BDE |
|---|---|---|---|---|---|
| *File System* | HDFS | HDFS | NFS | HDFS | **HDFS** |
| *Installation* | Native | Native | Native | Native | **lightweight virtualization** |
| *Flexible Modular Architecture* | no | no | no | no | **yes** |
| *High Availability* | Single failure recovery (yarn) | Single failure recovery (yarn) | Self healing, mult. failure rec. | Single failure recovery (yarn) | **Failure recovery** |
| *Cost* | Commercial | Commercial | Commercial | Free | **Free** |
| *Scaling* | Freemium | Freemium | Freemium | Free | **Free** |
| *Addition of custom components* | Not easy | No | No | No | **Yes** |
| *Integration testing* | yes | yes | yes | yes | -- |
| *Operating systems* | Linux | Linux | Linux | Linux | **Windows/Mac/Linux** |
| *Management tool* | Ambari | Cloudera manager | MapR Control system | - | **Docker swarm UI+ Custom** |

# BDE vs Hadoop distributions

- BDE is not built on top of existing distributions
- Targets
  - Communities
  - Research Institutions
- Bridges scientists and open data
- Multi Tier research efforts towards Smart Data

# Wrap Up, thanks for your Attention

Three possible streams for collaboration:

- A Big Data Platform for societal good
- Establishing data sharing and data value chains with the Industrial Data Space
- Semantic Domain Models (vocabularies, ontologies) for establishing a common understanding of the data

Please get in touch: Sören Auer (coordinator Big Data Europe), auer@cs.uni-bonn.de