Federated Analytics in the Dawn of the Digital Era
Federated Analytics in The Dawn of the Digital Era

Patricia Florissi, PH D
VP and Global CTO for Sales
There times in life where...
One must pause, think, and...
Try to grasp the moment we are living...
Many factors at play
Incremental computing reaching near zero marginal cost

Internet of things

Exponential technology advances every 5 years

Cloud

Big data analytics

Social

Mobile
Convergence of these Many factors at play
Dawn of a new era
Dawn of the digital era
What does this digital tide bring?
The things the tide brings

1. DATA SIZE MATTERS

2. DATA ECOSYSTEM MATTERS

3. DATA LOCALITY MATTERS
The things the tide brings

1. DATA SIZE MATTERS

2. DATA ECOSYSTEM MATTERS

3. DATA LOCALITY MATTERS
Cognitive revolution 2.0
Artificial intelligence

The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.
Cognitive revolution

- **DEEP LEARNING**
  - Example: MLPs

- **REPRESENTATION LEARNING**
  - Example: shallow autoencoders

- **MACHINE LEARNING**
  - Example: logistic regression

- **AI**
  - Example: knowledge bases
In the early days

Intellectually difficult for human being but relatively straight forward for computers

Deep Learning, Ian Goodfellow, Yoshua Bengio, and Aaron Courville

In the now days

Intellectually easy and intuitive for humans but hard to be described formally for computers
Machine learning

Learn from experience

Acquiring their own knowledge by extracting patterns from raw data
Deep learning

Learn from experience

Understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts.
Why is it called deep learning

- Represents the world as a nested hierarchy of concepts
- The depth of the representational graphs high
- Greater amount of composition of learned functions or learned concepts than traditional machine learning does
Learning algorithms

- Liner regression
- Logistic regression
- Support vector machines

Clustering
- K-means
- Hierarchical cluster analysis (HCA)
- Expectation maximization

Deep belief network (DBNs)

Visualization and dimensionality reduction
- Principal component analysis (PCA)
- Kernel PCA
- Locally-linear embedding (LLE)
- T-distributed stochastic neighbor embedding (T-SNE)

Polynomial regression
- K-nearest neighbors
- Anomaly detection
- Association rule learning
- Decision trees
- Random forests
- Ensemble methods
Common Denominator

They perform better

The more data

They analyze
The unreasonable effectiveness of data
Simple algorithms with lots of data will outperform sophisticated algorithms with less data.
1. DATA SIZE MATTERS

2. DATA ECOSYSTEM MATTERS

3. DATA LOCALITY MATTERS

The things the tide brings
The things the tide brings
Data monetization

The applicability of the data is large and broad.
Data fusion

Occurs when different sources are brought into contact and new facts emerge.
Digital world value chain

0101
1010
0110

1

Data: primary ingredient

2

Data: analytics: insights creation

3

Digital marketplace: insights creation

Applications: consumption vehicle
Digital world value chain graph
The things the tide brings
Major data analytics trends emerging
Centralization

And at the beginning there was one lake...

As the traditional approach to gain analytical insight from data in disparate sources

Batched data gathered in a single repository or cloud in preparation for analysis
The perfect data in-mobility storm

- IoT effect
- Bandwidth constraints
- Regulatory compliance
- Traditional Data Systems
- Multi-Cloud
Challenge imposed by traditional architecture thinking

Understanding the data challenge

**DATA HANDLING CHALLENGE**

- Massive numbers of endpoints for systems
- Storage requirement is unknown
- Managing massive, heterogeneous data an enormous challenge.

Diverse data sources -> diverse protocols, diverse transport requirements, diverse data

Must normalize and standardize for future data orchestration and integration

**DATA USAGE CHALLENGE**

Need to create an innovative business that uses this data

->

creates more value and fully uses current investment
Where is the lake?
The things the tide brings
The things the tide brings...

1. Data Size Matters
   - Incremental computing reaching near zero marginal cost
   - Exponential technology advances every 5 years

2. Data Ecosystem Matters
   - Cloud
   - Big data analytics
   - Social

3. Data Locality Matters
   - Mobile
In the dawn of the digital era ...
The dawn of federated analytics
Federated analytics

Distributed analysis of data in place, as close as possible to the data source, while intermediate results can be shared and can be further analyzed along their path to other locations, enabling higher order learning at scale.
Where is the lake?
Centralization

And at the beginning there was one lake...

As the traditional approach to gain analytical insight from data in disparate sources

Batched data gathered in a single repository or cloud in preparation for analysis
There is a multi-lake
Anatomy of the world's data
Anatomy of a virtual data fabric
Main challenges

*how* data science algorithms will be re-designed to reason and learn in a federated manner;

*how* analytics will be distributed close to where the data is collected; and

*how* the intermediate results will be aggregated and analyzed together to drive higher-order learning at scale. Those are much more challenging problems to solve.
World Wide Herd (WWH)
Data sources generate data
Along comes the analytics explorer
Parallel and distributed computation
Parallel and distributed computation

Virtual Computing Cluster

Analytics Fabric

Data-zone Fabric

Data-zones
Privacy-preserving results shared

Virtual Computing Cluster

Analytics Fabric

Data-zone Fabric

Virtual Computing Nodes

Data-zones
Broad Range of Data-zones
IoT gateways

Virtual Computing Cluster

Analytics Fabric

Data-zone Fabric

Virtual Computing Nodes

Data-zones
HCI receiving data from digital buildings

Virtual Computing Cluster

Analytics Fabric

Data-zone Fabric

Virtual Computing Nodes
HCI receiving data from sensors

Virtual Computing Cluster

Analytics Fabric

Data-zone Fabric

Privacy preserving results

Virtual Computing Nodes
Cloud instances receiving data from sensors

Analytics Fabric

Virtual Computing Cluster

Virtual Computing Nodes

Data-zone Fabric

Privacy preserving results

Data-zones
Cloud Instances receiving data from Digital Buildings

Virtual Computing Cluster

Analytics Fabric

Virtual Computing Nodes

Data-zone Fabric

Privacy preserving results

Data-zones
Computing nodes become virtual nodes in a virtual computing cluster at scale
Connecting clusters at the analytics layer
Rebirth of data science
Benchmarking
Parameters for Histogram & Median Use Case

- Attribute selected: age, ranging form 20 to 80
- Number of records: 275 total
- Three data-zones:
  - purple,
  - blue, and
  - gray
Summary of Results

- **Main Cluster**: Percentile value: 43 (2nd slice)
- **Remote cluster 1**: Percentile value: 54 (3rd slice)
- **Remote cluster 2**: Percentile value: 63 (4th slice)
- **Global Percentile Range**: 55 – 56 (4th slice)
WWH Fabrics

Because “the future is federated”
Federate analytics per geography
How all these connections get established?
Yet another fabric
The meta-data fabric

Virtual Computing Cluster

Analytics Fabric

Meta-data Fabric

Data-zone Fabric

Data-zones
Catalog nodes

Virtual Computing Cluster

Analytics Fabric

Meta-data Fabric

Data-zone Fabric

Catalog Nodes

Data-zones

Virtual Computing Nodes
Persona

A specific, fictional representation of a person who is very likely to represent your ideal customer. A persona is a prototype of your user.

https://content.pivotal.io/blog/how-to-define-your-persona-part-2
Business value creator

Virtual Computing Cluster

Analytics Fabric

Data Scientist

Virtual Computing Nodes
Business value harvester

Virtual Computing Cluster

Analytics Fabric

Analytics Explorer
Continuous feedback loop

Virtual Computing Cluster

Analytics Fabric

Virtual Computing Nodes

Analytics Explorer
Analytics Explorer
Analytics Explorer
Data Scientist
Data Scientist
Data Scientist
Together, business transformation happens

Analytics Fabric

Virtual Computing Cluster

Virtual Computing Nodes

Analytics Explorer

Data Scientist
Analytics requires data and processing capacity
Data architects manage data

Data Architect

Data-zone Fabric

Data-zones
Administrators provision processing nodes
Together, data and processing capacity are created
Different data-zones, different set of owners

Data-zone Fabric

Data-zones
Meta-data curator catalogs it all
Creating the meta-data fabric
Computing nodes become virtual nodes in a virtual computing cluster at scale
Some people make it happen
Some people watch it happen
Some people do not know what happened
Think bigger data…

Think distributed bigger data…

Think world wide scale, distributed bigger data…
Dawn of a new era
Technology leadership is a team sport
We are better together
Better together because…

only together we can do things no one else can
The time is now
To change the world
...one challenge at a time
Let your transformation begin!
Thank you!