Big Data Analytics & IoT

Instructor: Ekpe Okorafor

1. Accenture – Big Data Academy
2. Computer Science - African University of Science & Technology

Ekpe Okorafor PhD

Affiliations:

- **Accenture Digital – Big Data Academy**
  - ❑ Senior Principal & Faculty, Big Data & Analytics
- **African University of Science & Technology**
  - ❑ Visiting Professor, Computer Science / Data Science
  - ❑ Research Professor - High Performance Computing Center of Excellence

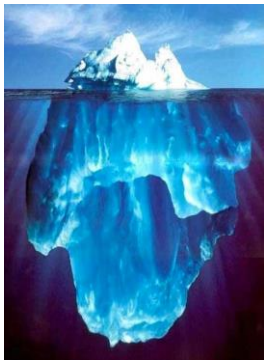
Research Interests:

- Big Data, Predictive & Adaptive Analytics
- Artificial Intelligence, Machine Learning
- Performance Modelling and Analysis
- Information Assurance and Cybersecurity.
- High Performance Computing & Network Architectures
- Distributed Storage & Processing
- Massively Parallel Processing & Programming
- Fault-tolerant Systems

Email: ekpe.okorafor@gmail.com; eokorafo@ictp.it; eokorafor@aust.edu.ng
Twitter: @EkpeOkorafor; @Radicube

Where does data come from?

It's All Happening On-line

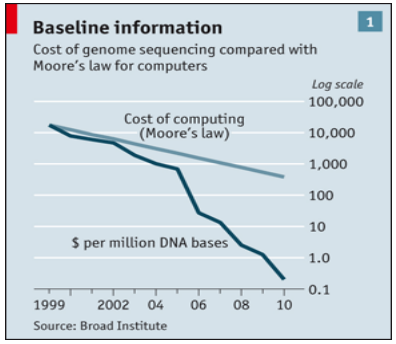


Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

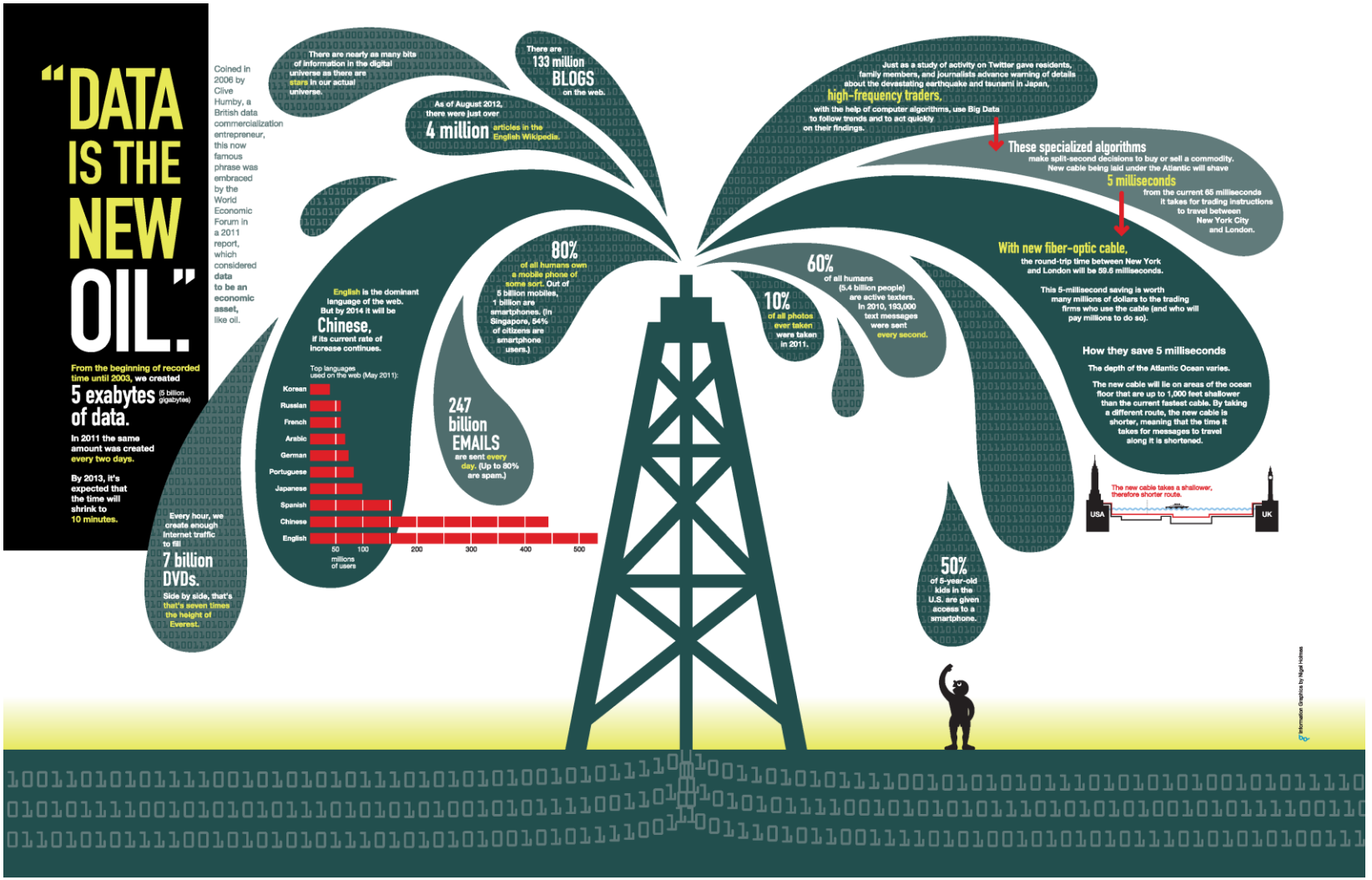
User Generated (Web & Mobile)

Internet of Things / M2M

Health/Scientific Computing

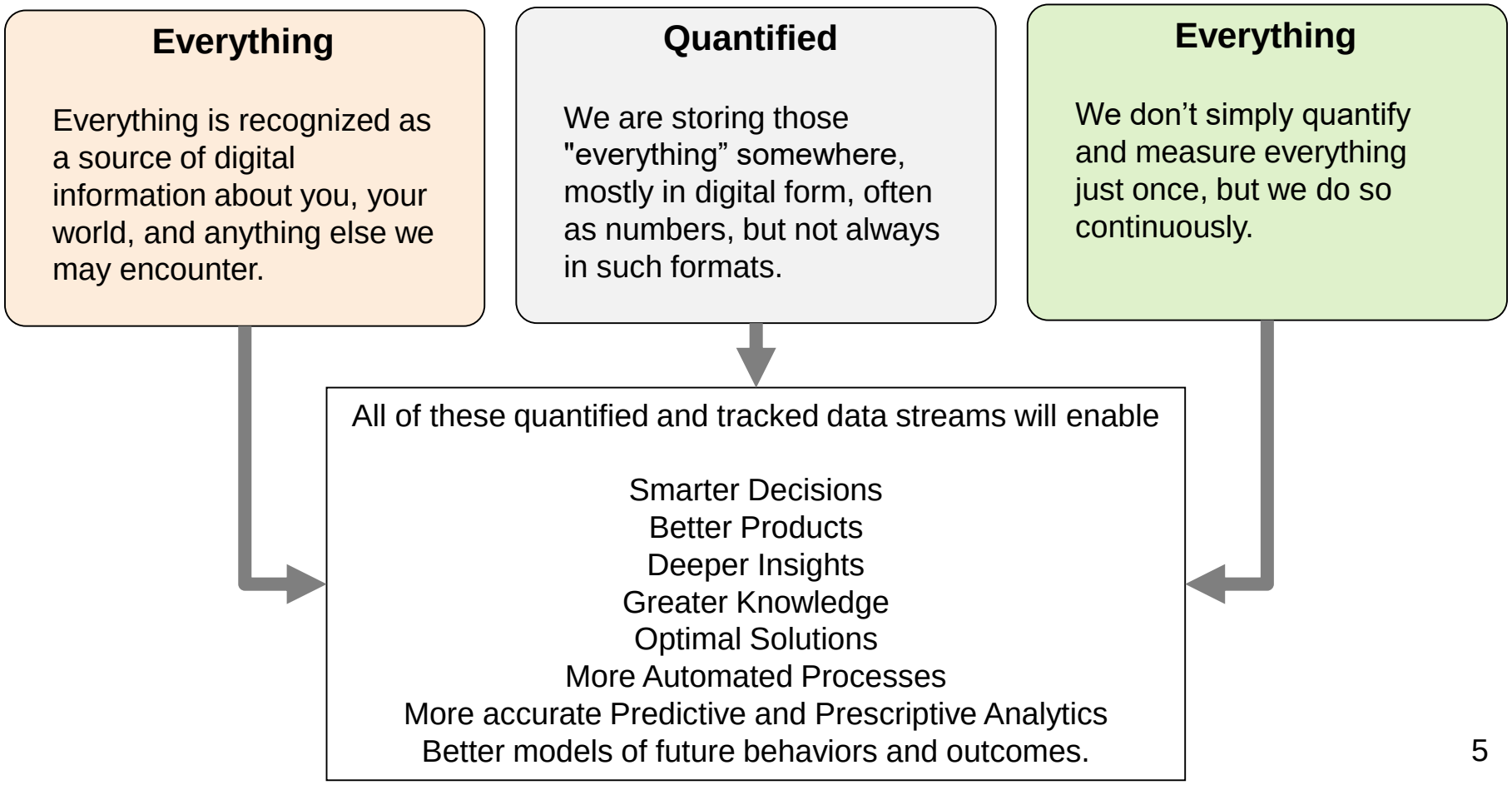


“Data is the New Oil” – World Economic Forum 2011



What is Big Data?

According to the Author Dr. Kirk Borne, Principal Data Scientist, Big Data Definition is **Everything, Quantified and Tracked.**

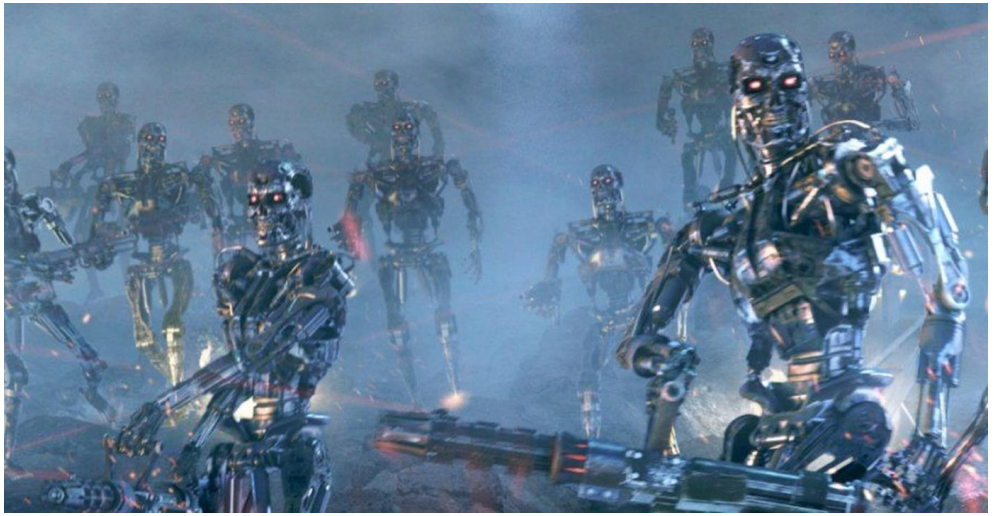


What is IoT?

The Internet of Things (IoT) is the network of physical objects—devices, vehicles, buildings and other items embedded with electronics, software, sensors, and network connectivity—that enables these objects to collect and exchange data.

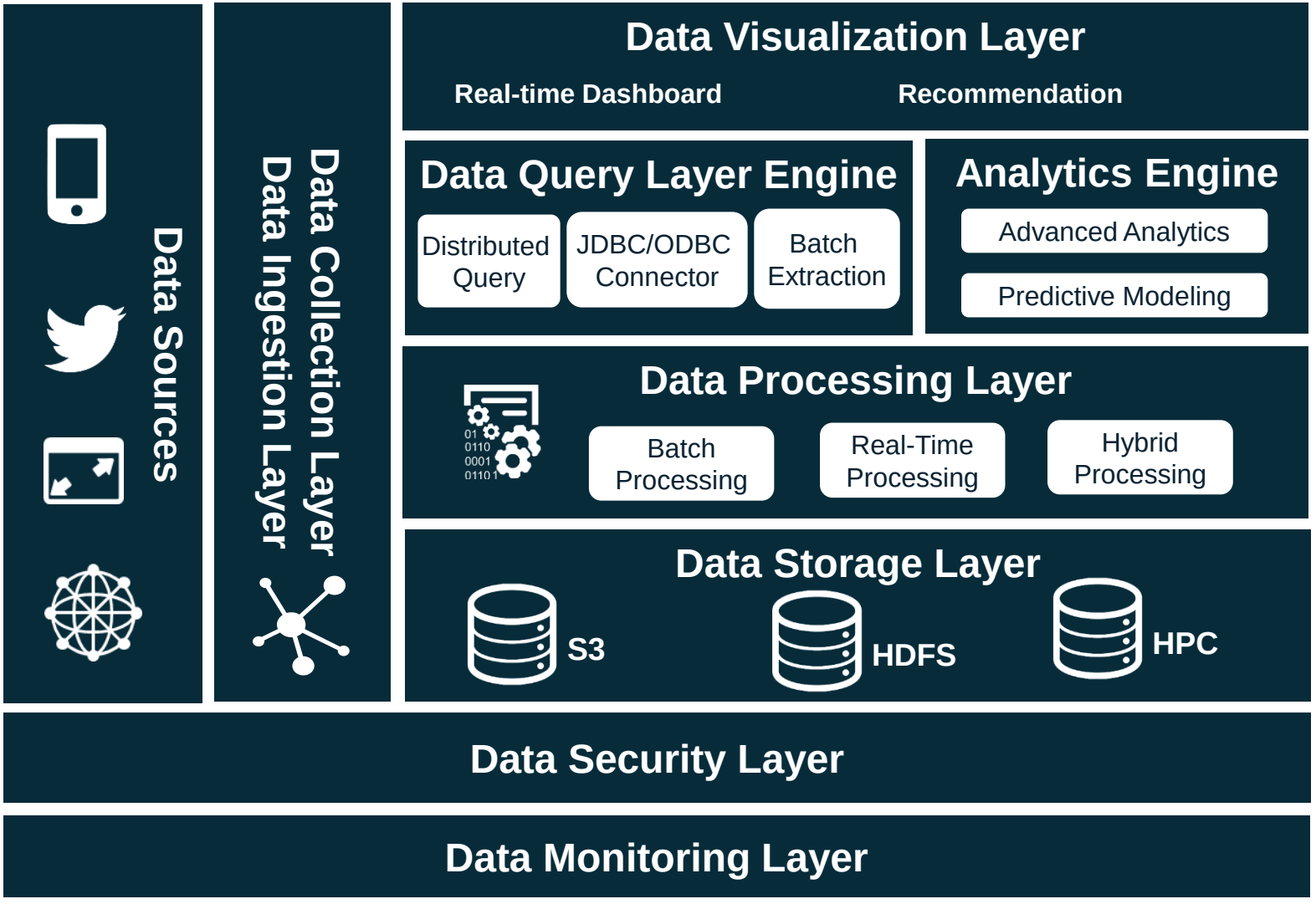
Various Names, One Concept

- M2M (Machine to Machine)
- “Internet of Everything” (Cisco Systems)
- “World Size Web” (Bruce Schneier)
- “Skynet” (Terminator movie)



Connect. Compute. Communicate.

How Do We Handle Big Data / IoT? - Big Data Framework



Data Ingestion Layer

Big Data Ingestion involves connecting to various data sources, extracting the data, and detecting the changed data. It's about moving data - and especially the unstructured data - from where it is originated, into a system where it can be stored and analyzed.

- Challenges: with IoT, volume and variance of data sources
- Parameters: velocity, size, frequency, formats
- Key principles: network bandwidth, right tools, streaming data
- Tools: Apache Flumes, Apache Nifi

Data Collection (Integration) Layer

In this Layer, more focus is on transportation data from ingestion layer to rest of Data Pipeline. Here we use a messaging system that will act as a mediator between all the programs that can send and receive messages.

- Kafka works with Storm, Hbase, Spark for real-time analysis and rendering streaming data
  - Building Real-Time streaming Data Pipelines that reliably get data between systems or applications
  - Building Real-Time streaming applications that transform or react to the streams of data.
- Data Pipeline is the main component of data integration

Data Processing Layer

In this Layer, data collected in the previous layer is processed and made ready to route to different destinations.

- **Batch processing system** - A pure batch processing system for offline analytics (Sqoop).
- **Near real time processing system** - A pure online processing system for on-line analytic (Storm).
- **In-memory processing engine** - Efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets (Spark)
- **Distributed stream processing** - Provides results that are accurate, even in the case of out-of-order or late-arriving data (Flink)

Data Storage Layer

Next, the major issue is to keep data in the right place based on usage. A combination of distributed file systems and NoSQL databases provide scalable data storage platforms for Big Data / IoT

- **HDFS** - A Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers.
- **Amazon Simple Storage Service (Amazon S3)** - Object storage with a simple web service interface to store and retrieve any amount of data from anywhere on the web.
- **NoSQL** – Non-relational databases that provide a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases.

Data Query (Access) Layer

This is the layer where strong analytic processing takes place. Data analytics is an essential step which solved the inefficiencies of traditional data platforms to handle large amounts of data related to interactive queries, ETL, storage and processing

- **Tools** – Hive, Spark SQL, Presto, Redshift
- **Data Warehouse** - Centralized repository that stores data from multiple information sources and transforms them into a common, multidimensional data model for efficient querying and analysis.
- **Data Lake** - Cloud-based enterprise architecture that structures data in a more scalable way that makes it easier to experiment with it. All data is retained

Data Visualization Layer

Real-time Dashboard **Recommendations**

This layer focus on Big Data Visualization. We need something that will grab people's attention, pull them in, make your findings well-understood. This is the where the data value is perceived by the user.

- **Dashboards** – Save, share, and communicate insights. It helps users generate questions by revealing the depth, range, and content of their data stores.
  - Tools - Tableau, AngularJS, Kibana, React.js
- **Recommenders** - Recommender systems focus on the task of information filtering, which deals with the delivery of items selected from a large collection that the user is likely to find interesting or useful.