Privacy Enabled Big Data Analytics in the Cloud

Humboldt Colloquium – Tel-Aviv September 2016

Shelly Garion
IBM Research - Haifa
My German Experience
Big Data Challenges

- Where to store it?
- How to analyze it?
- What about privacy?
Cloud Object Store

Object

Data
(unstructured / semi-structured)

Metadata

AWS Blog
Amazon S3 – Two Trillion Objects, 1.1 Million Requests / Second
by Jeff Barr on 18 April 2013 in Amazon S3 | Permalink | Comments
Apache Spark

Fast and general open-source engine for large-scale data processing

- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk
- Write applications quickly in Scala, Python, Java, or R
- Combines SQL, streaming and complex analytics (machine learning, graph processing)
- Runs on Apache Mesos, Hadoop YARN cluster manager, standalone, or in the cloud. It can access diverse data sources (HDFS, S3, Openstack Swift Object Store)
- Built by a wide set of developers from over 200 companies. Since 2009, more than 1000 developers have contributed to Spark

https://spark.apache.org/
Privacy Concepts

From European Union Data Protection Directive

• **Personal data** – “any information relating to an identified or identifiable **natural person**”

• Data subject (person) has the right to be informed when his personal data is being processed

• **Consent** – Data may be processed only when the data subject has given his consent

• **Proportionality** – The data processed and the time for which it is stored should be no more than required for the stated **purpose**

Data Security

“Protecting data from destructive forces and from the unwanted actions of unauthorized users.”

Wikipedia

Data Privacy

Ensuring that personal data is used and stored only as is needed to provide the approved user services.
Privacy Enabled Analytics Architecture

REST API

Consent Manager Service

Home owner consent GUI

Personal Recommendations

Crowd-based Recommendations

Analytics use-cases:
- Heating schedule
- Damp prediction

Data Access Controller

Filtered data

Smart house: temperature, humidity, window sensors

Analytics Engine

Object storage

Consent Manager Service

Data Access Controller

REST API