Cloud computing based big data ecosystem and requirements

Yongshun Cai (蔡永顺)
Associate Rapporteur of ITU-T SG13 Q17
China Telecom

Dong Wang (王东)
Rapporteur of ITU-T SG13 Q18
ZTE Corporation
Agenda

• ITU-T progress with big data
• Big data characteristics and challenges
• Cloud computing based big data ecosystem
• Use cases of big data applications & services
• Requirements and capabilities of cloud computing based big data
• Conclusion
ITU-T progress with big data

- **2013, July:**
  - Initiation of 1st big data working item Y.Bigdata-reqts (Requirements and capabilities for cloud computing based big data) by ITU-T SG13 Q17
    - Overview of cloud computing based big data;
    - Big Data system context and its activities;
    - Cloud computing based big data requirements, capabilities, use cases and scenarios
  - Targeted to be consented in July 2015

- **2013, November:**
  - ITU-T watch report “Big Data: Big today, normal tomorrow” by TSAG

- **2014, July:**
  - Initiation of 2nd big data working item Y.IoT-BigData-reqts (Specific requirements and capabilities of the Internet of Things for Big Data) by ITU-T SG13 Q2

- **2015, April:**
  - Initiation of further big data working items:
    - Y.BigDataEX-reqts (Big Data Exchange Requirements and Framework for Telecom Big data))
    - Y.Suppl.BigData-RoadMap (Supplement on Big Data Standardization Roadmap)
    - Y.BDaaS-arch ((Functional architecture of Big Data as a Service)
Big data definition and characteristics

Definition
- **Big data**: a category of technologies and services where the capabilities provided to collect, store, search, share, analyse and visualize data which have the characteristics of high-volume, high-velocity and high-variety.

Characteristics
- **Volume (large)**: refers to the amount of data collected, stored, analyzed and visualized, which need big data technologies to be resolved.
- **Variety (Diverse)**: refers to different data types and data formats that are performed by big data applications and platforms.
- **Velocity (High)**: refers to both how fast the data is being collected and how fast the data is processed to deliver expected results.
Big data challenges

- **Heterogeneity and incompleteness**
  Data processed using big data technologies can miss some attributes or introduce noise in data transmission. Even after data cleaning and error correction, some incompleteness and errors in data are likely to remain.

- **Scale**
  Managing large and rapidly increasing volumes of data is a challenging issue for data processing. In the past, the data scale challenge was mitigated by evolution of processing and storage resources.

- **Timeliness**
  The acquisition rate and timeliness, to effectively find elements within limited time that meet a specified criterion in a large data set, are new challenges faced by data processing.

- **Privacy**
  Data about human individuals, such as demographic information, internet activities, commutation patterns, social interactions, energy or water consumption, are being collected and analysed for different purposes.
Cloud computing based big data ecosystem

The roles in cloud computing can support the three main roles of big data as followings:

- CSN: data provider;
- CSP: big data application provider;
- CSP: big data infrastructure provider;
- CSC: big data service customer.
Use cases 1- Personalization customized service

- Each web access activity might be remains as logs of records of visiting Web sites
- The CSP:BDIP store relevant information of the Web service user’s activities
- The CSC:BDAP may know the patterns or preference of the user through analysis of the user’s activities
Use cases 2- Intelligent transport big data analysis

- The traffic of vehicles and the status of roads are collected in real time. The big data applications for intelligent transport systems are built.
- Real time and accurate route navigation
- Accurate scheduling for bus departures
- Automatic recognition of violations of laws on road traffic
Requirements & capabilities 1– Data collection

- **Data source recognition**, which offers the capabilities to locate the data sources and detect the types of data being exposed;
- **Data adaptation**, which offers the capabilities to transform and organize the data being accepted with targeted structured and attributes (numbering, location, ownerships, etc);
- **Data transmission**, which offers the capabilities to transfer data sets from one location to another keeping the integrity and consistency.
- **Data extraction**, which offers the capabilities to extract the semi-structured data and unstructured data;
- **Data import**, which offers the capabilities to import large amount of static data and real-time data;
- **Data integration**, which offers the capabilities to integrate data from different data sources (different data type or schema) using by metadata or ontology;
Requirements & capabilities 2– Data storage

- **Data registration**, which offers the capabilities to create, update and delete the metadata with corresponding changes to data storage;
- Note - In case of non-structured data registration, data registration component can request the transforming raw data to semi-structured data (e.g. JSON, BSON) for data registration;
- **Data persistence**, which offers the capabilities to store data with low redundancy and low cost and converged with processing capability;
- **Data semantic intellectualization**, which offers the capabilities to define semantic relationships among different data set for knowledge sharing;
- **Data access**, which offers the capabilities to access data through multiple interfaces, such as API;
- **Data indexing**, which offers the capabilities to generate and update index for data sets;
- **Data duplication and backup**, which offers the capabilities to duplicate and make backup for data sets;
Requirements & capabilities 3– Data analysis

- **Data preparation**, which offers the capabilities to transform data into a form that can be analyzed. This capability includes exploring, de-noising, changing, and shaping the data;
- **Data analysis**, which offers the capabilities to investigate, inspect, and model data in order to discover useful information;
- **Data visualization**, which offers the capabilities to display and present data analysis results;
- **Workflow automation**, which offers the automation processes, in whole or part, during which data or functions are passed from one step to another for action, according to a set of procedural rules;
- **Analysis algorithm adaption**, which offers the capabilities to apply algorithms of classification, regression, clustering, association rules, ranking and so on according to the requirements;
Requirements & capabilities 4– Data management

- **Data provenance**, which offers the capabilities to manage information pertaining to any source of data including the party or parties involved in generation, introduction and/or mash-up processes for data;
- **Data preservation**, which offers the capabilities to manage the series of activities necessary to ensure continued access to data for as long as necessary;
- **Data privacy**, which offers the capabilities to manage the right of individuals to control or influence what information related to them;
- **Data security**, which offers the capabilities to handle the network and service aspects of security, including administrative, operational and maintenance issues;
- **Data ownership**, which offers the capabilities to manage digital right of data possession, disposition according to the change of data status (e.g. data integration);
- **Processing task track and control**, which offers the information such as success of the job and task, running time and resource utilization and etc.
Conclusion

Y.Bigdata-reqts “Requirements and capabilities for cloud computing based big data”

- This Recommendation is expected to be delivered in Q3 2015

- User view approach:
  - Big data system context with activities
  - Cloud computing based big data ecosystem and activities

- Usecase to functions derivation:
  - Use cases of cloud computing in support of big data
  - Use cases of cloud computing based big data as analysis services (BDaaS)

- Main requirements / capabilities cover the whole process of big data processing:
  - Data collection
  - Data storage/persistence
  - Data analyse
  - Data visualization/application
  - Data management/security/privacy

- As a basis or reference for other big data recommendations
  - Y.BDaaS-arch
  - Y.IoT-BigData-reqts
  - Y.BigDataEX-reqts
  - .....
Meeting plan in 2015

- July 13-23, Geneva, Rapporteur meeting
- Nov 30–Dec 11, Geneva, Plenary meeting

Welcome to join us

Q&A

Thank you for your attention