FUJITSU

shaping tomorrow with you

Fujitsu/Fujitsu Labs' Technologies for Big Data in Cloud and Business Opportunities

Satoshi Tsuchiya
Cloud Computing Research Center
Fujitsu Laboratories Ltd.
January, 2012

Overview: Fujitsu's Cloud and Big Data

FUJITSU

- Fujitsu IaaS FGCP/S5: already deployed in world wide
  - Public IaaS cloud platform
  - Beta started in 2009, now deployed in 5 locations worldwide
  - Pay for what you use / Elastic and scalable
- Fujitsu's PaaS for Big Data Convergence Services Platform (planned)
  - PaaS for "Big Data": Integrated Environment
    event processing, parallel batch(MapReduce), etc.
  - Announced in Aug. 2011 /
    Early Beta service will start in March, 2012 (in Japan)

Cloud Computing Research Center at Fujitsu Labs is working on R&D of key technologies for Fujitsu Cloud Services.

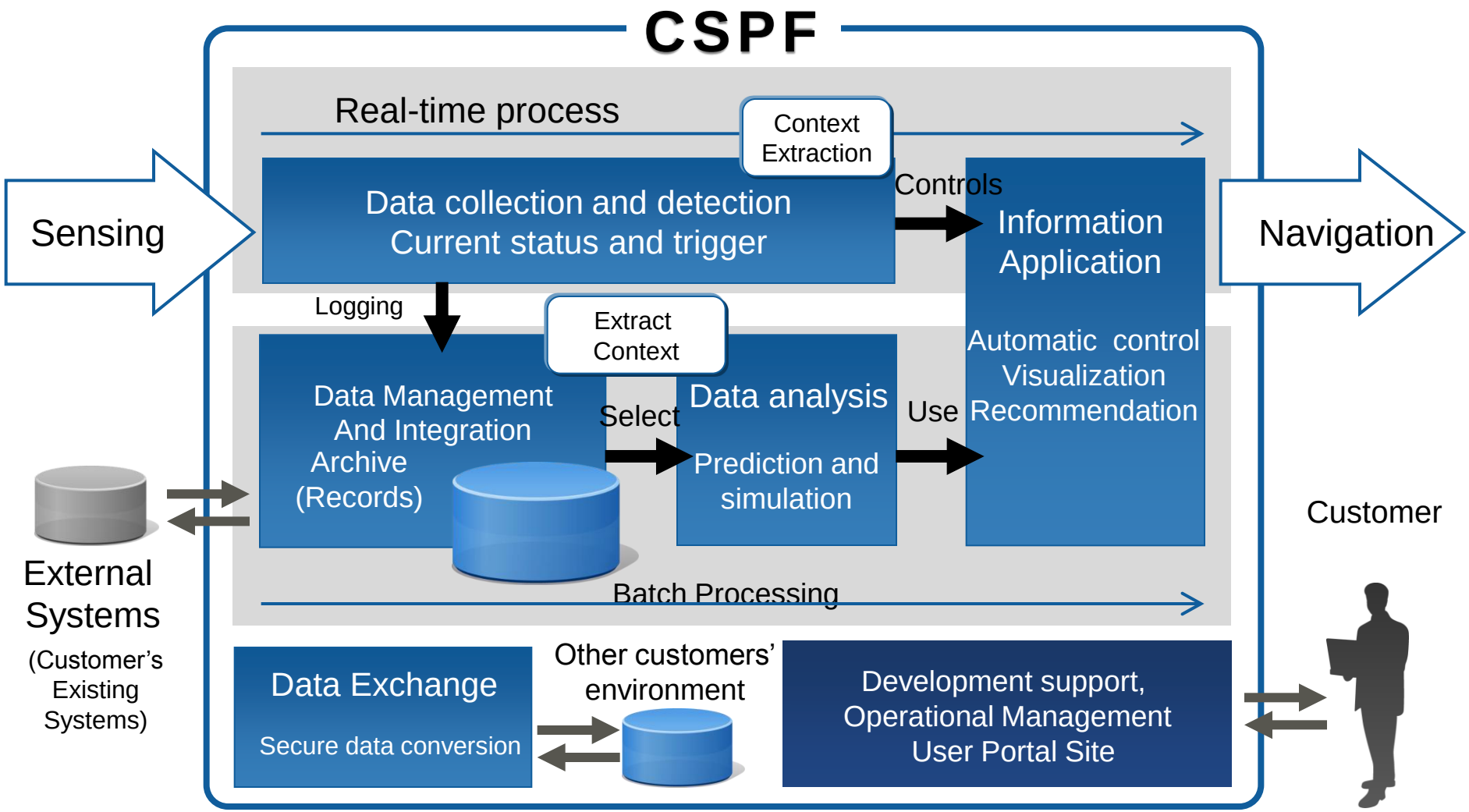
My research team focuses on "Parallel Data Processing".

Convergence Services Platform (PaaS)

FUJITSU

Integrated, easy-to-use data processing functions on the Fujitsu Cloud
Announced August 2011, early beta service will start from March 2012

http://www.fujitsu.com/global/news/pr/archives/month/2011/20110830-01.html



New Challenges on Big Data

FUJITSU

- Gartner: 3 challenges on Big Data (June 2011)
  - Volume: store enormous amount of data (tens of TB ~ several PB)
  - Variety: transaction logs, sensor records, image, video, etc.
  - Velocity: competitiveness depends on the responsiveness of analysis

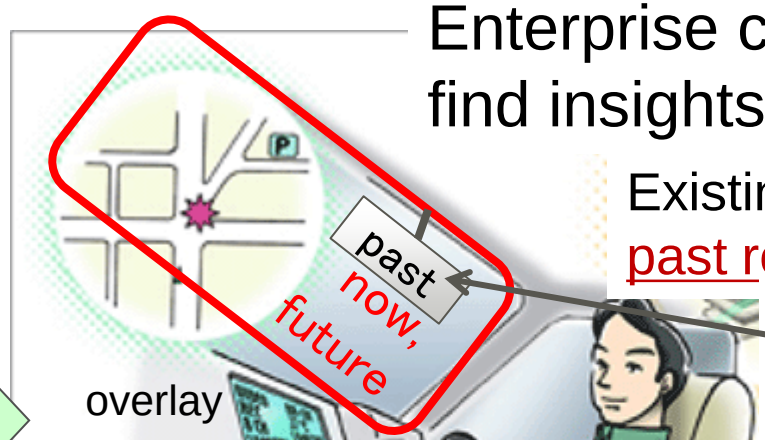
  ➔ Not just Volume, **Volume and Velocity together**

- Advanced Users needs Velocity in tens of TB
  - The report “Big Data Analysis” (Data Warehousing Institute)
    - Many advancing analysis users want to get results within hourly (min ~ sec)
      (Those advanced users already have tens of TB)

Big Data is like driving a car in the sea of information

FUJITSU

Enterprise customers wants to find insights from the real world.



Existing IT systems only shows past results in a small window

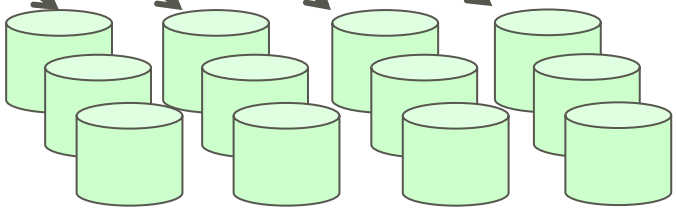
Record of past

relatively small

The Real World

ever-changing

ever-growing



various sources, enormous amount

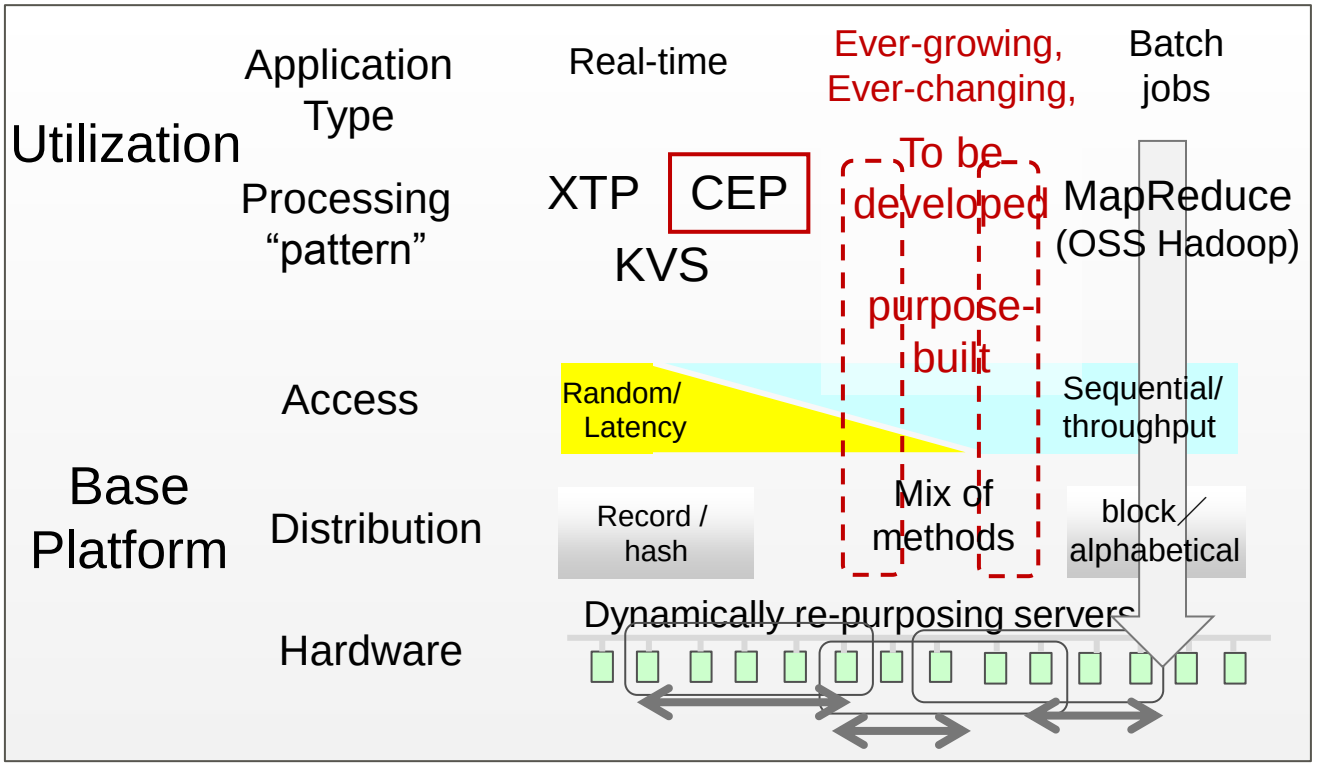
New IT systems expected to show
- Now : visualize current situation
- Future: prediction, recommendation

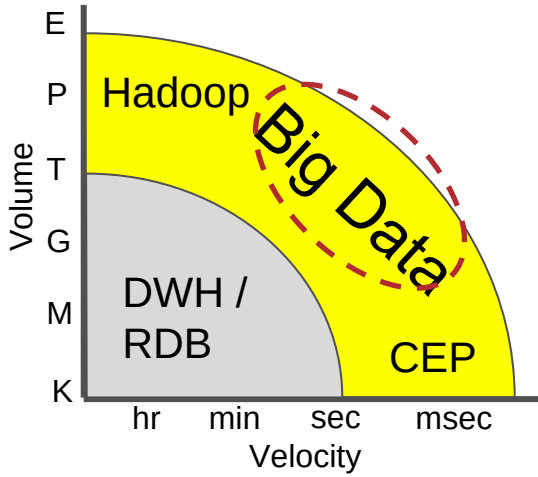
from enormous and ever-changing, ever-growing data

The Technology Map of Big Data Processing

FUJITSU

There is no single ring to rule them all.





XTP: eXtream Transaction Processing

CEP: Complex Event Processing

KVS: Key-Value data Store

- Two major purpose-built towers: Real-time and Batch in parallel
  - Real-Time: **Latency focused** ➔ record-base, short msgs, allocation by hash (random acc)
  - Batch in parallel: **Throughput focused** ➔ Big block in storage, sequential/sorted allocation
- Next Step: variety of purpose-built systems
  - mix of methods/elements appropriately for each need of enterprises

Exhibit

A highly parallel and fast range query function for a distributed data store

FUJITSU

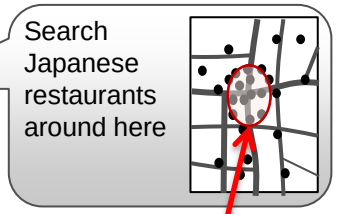
"**Distributed KVS (Key-Value Store)**" provides a storage function with scalability and fault tolerancy.

However

A rich function like "**Range Query**" cannot be executed efficiently on existing distributed KVS tech.

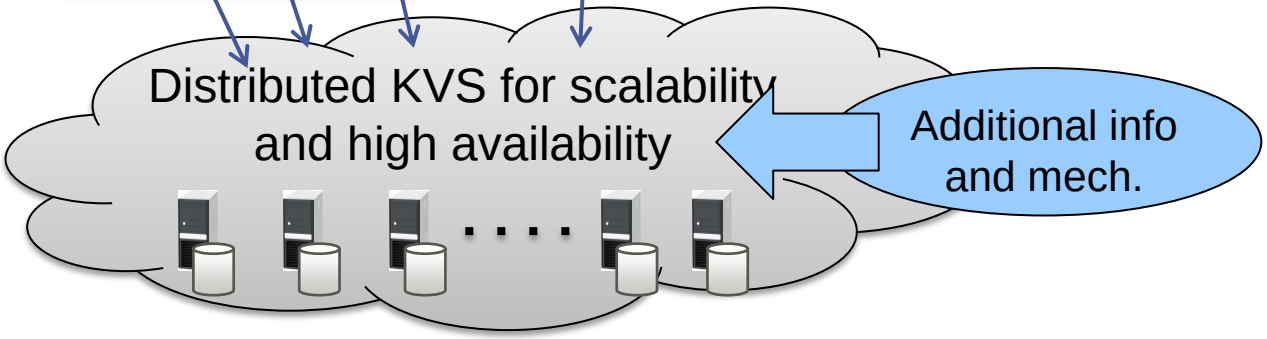
Multitude of Sensors

Various functional Services



"**Range Query**" is a data extraction technique from a data set

Data Accumulation (24 hours 365 days)

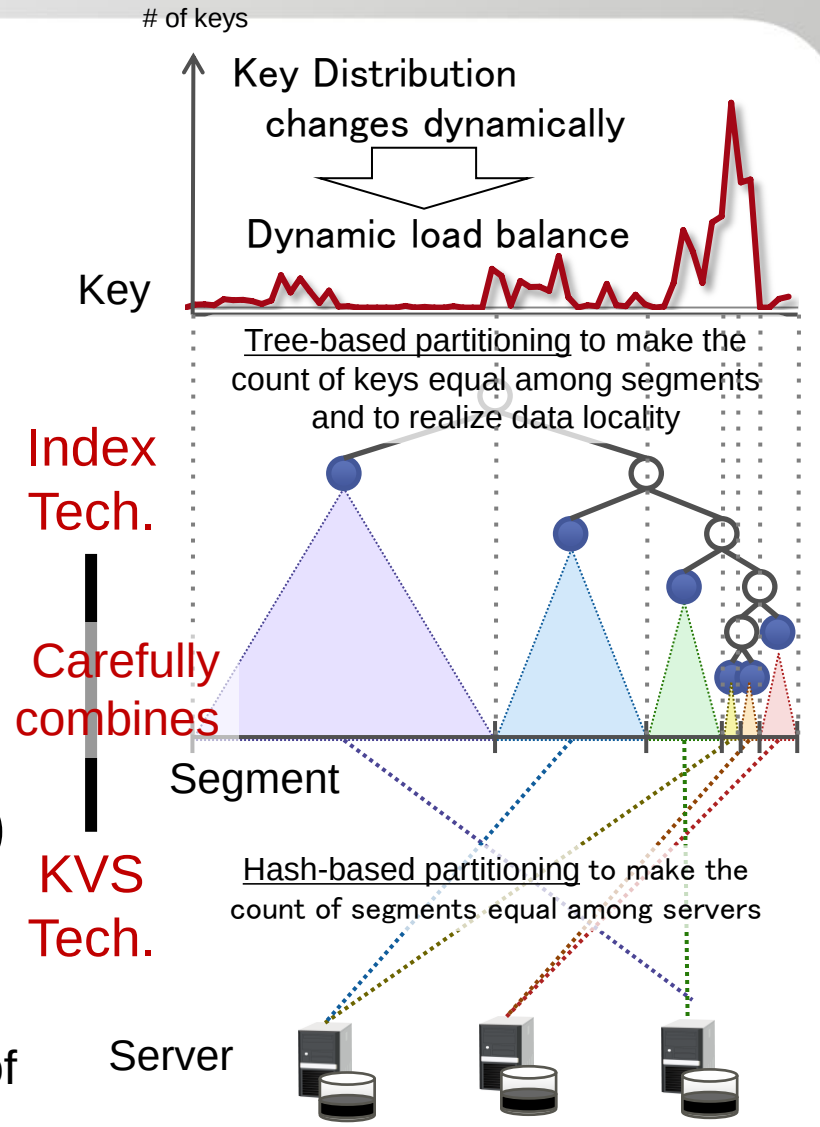


Range Query needs additional info. and mechanism for rapid and efficient response

- No Index (a simple answer) query to all possible nodes
  → Very Inefficient
- Centrally managed Range Index ex. Hbase (Hadoop KVS)
  → bad at scale out operation it needs careful design

Technology Enablers

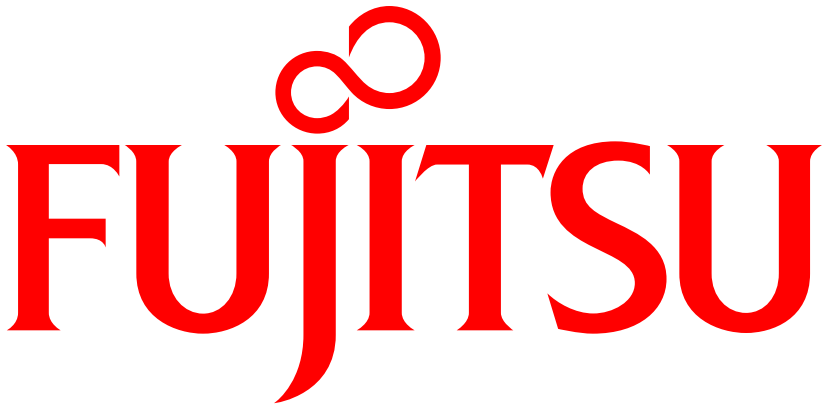
- Two-layer data partitioning technique and combines them careffully in a distributed manner
  - key ←→ segment (for efficiency)
    - Put keys close to each other into the same segment (locality-aware)
    - Tree-based allocation
    - Dynamically split segments based on the accumulated amount of data (load balancing in terms of volume)
  - segment ←→ server (for high avail.)
    - Put segments into servers randomly
    - Hash-based allocation
    - Preserve high availability and scalability of distributed KVS



Summary

FUJITSU

- Big Data is not just for Volume, **Volume and Velocity together**
- Big Data is like driving a car in the sea of information
  - Existing IT system treats relatively small data and just show the past trends in a small rear view window.
  - New IT systems are expected to show the future (prediction, recommendation) in a big front window (for rapid, precise decision)
- Next phase is variety of purpose-built systems to fulfill specific enterprise needs
  - Basic data processing functions (Event Processing / Parallel Batch) are available
  - Mix of methods/elements to fulfill the requirements of each enterprise with understanding elemental tech. and carefully designed combinations
- Fujitsu Labs are developing high level functions on top of basic parallel technologies aiming at purpose-built Big Data system in the cloud.



shaping tomorrow with you