Big Data in Enterprise
challenges & opportunities

Yuanhao Sun
孙元浩
yuanhao.sun@intel.com
Software and Service Group

Big Data Phenomenon

1.8ZB in 2011

2 Days > the dawn of civilization to 2003

750M

Photos uploaded to Facebook in 2 days

966PB

Stored in US manufacturin (2009)

20TB/hour

Sensor output of a Boeing jet engine

200+TB

A boy's 240'000 hours by a MIT Media Lab geek

200PB

Created by a Smart City project in China

$800B

in personal location data within 10 years

$300B/year

US healthcare saving from Big Data

$20+B

Acquisitions in the last 12 months

Data are becoming the *new raw material of business*: an economic input almost on a par with capital and labor.

The Economist, 2010

Information will be the "*oil of the 21st century*".

Gartner, 2010

intel Software

intel

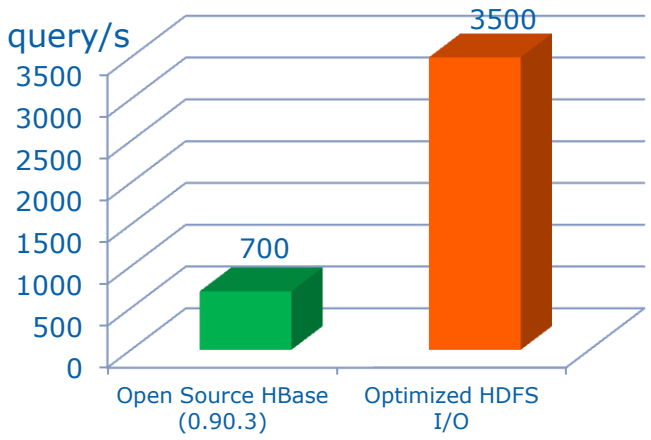
Big Data in Telecom

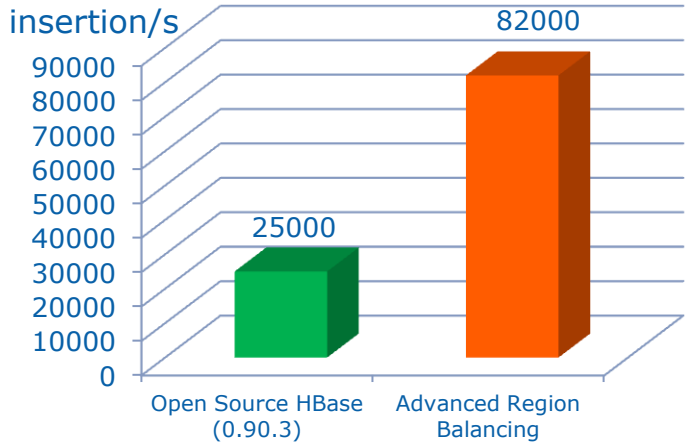
- Lots of data
  - One telco operator: 360TB Call Data Records within 6months (in a provincial branch, 100M users)
  - The other operator: ~300TB web access logs from mobile phones within 6 months
- Keep growing
  - ~2TB CDR/day in a provin
- Various dat
  - CDR, GPRS, 3G, WLAN, Value-add services, etc)
  - Billing data, accounting data, sales & marketing data, etc.
  - Web access logs
  - Network signaling data
  - Base station sensor data

Difficult to manage and monetize these data!

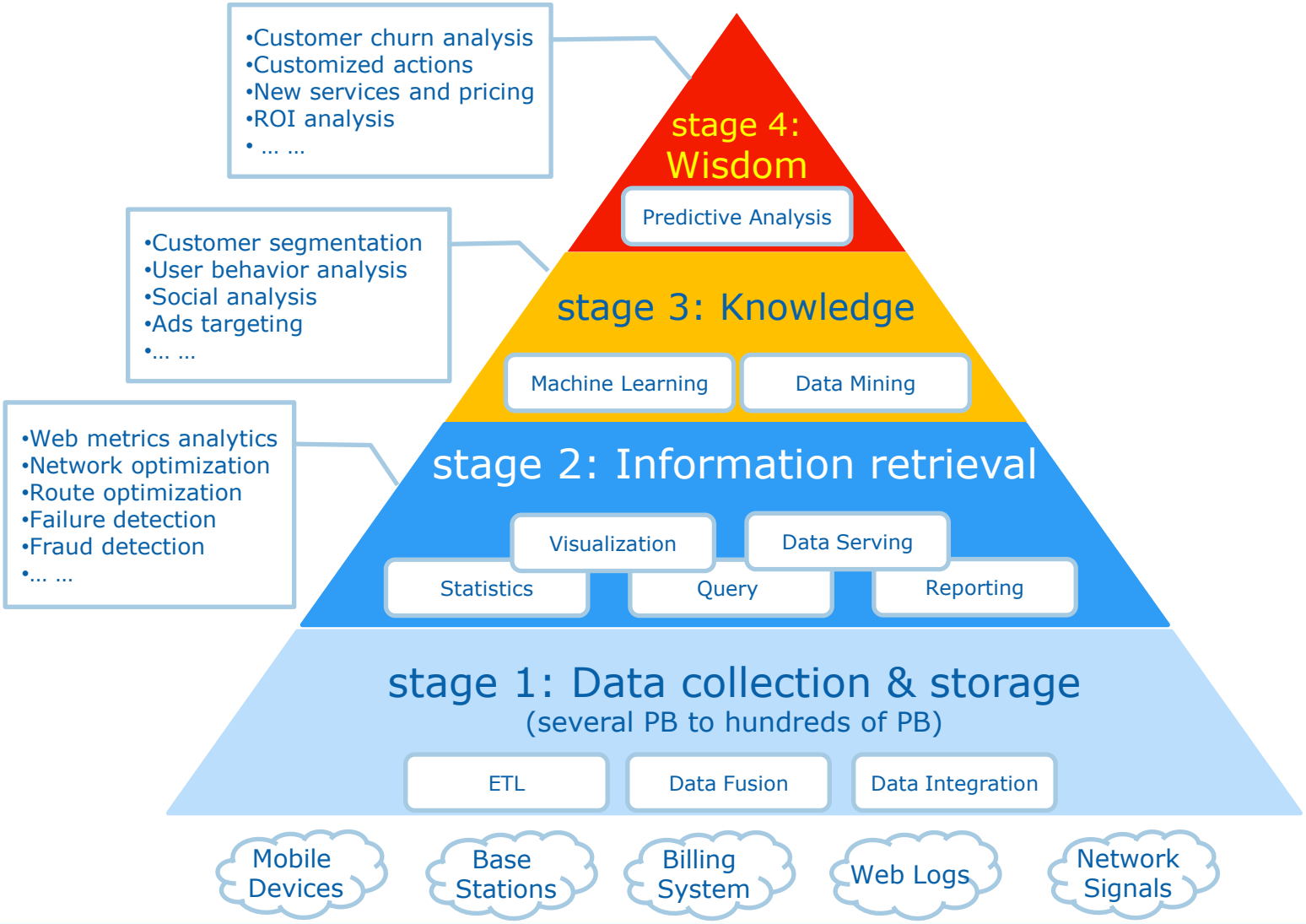
How Hadoop helps

- Map/Reduce for data loading and data cleansing
- HBase as the data store
  - Inserting 10000 records/second/server (2-way, 32GB) in average
  - Read from disk: >400 query/second/server, latency within one second (0.05s~0.8s under different load)
    - A query is a scan to get all CDR within one month for one user.
- Optimizations significantly increase the throughput of a 8-node cluster





Value chain of big data in telecom



Healthcare: Care Coordination and Data Sharing for Improved Outcomes

Quality of Life

Proactive health and Wellness

Reduce illness. Promote wellness and empowerment

Home Care

Reduce costly emergency care. Better manage chronic disease.

Residential/Community /Ambulatory Care

Reduce hospital (re) admissions. Manage at home.

Acute Care

Reduce ALOS. Earlier Discharge to Ambulatory environments.

Cost of Care

Highest Quality of Life at the Lowest Possible Cost

Enabling Technologies for Coordinated Care

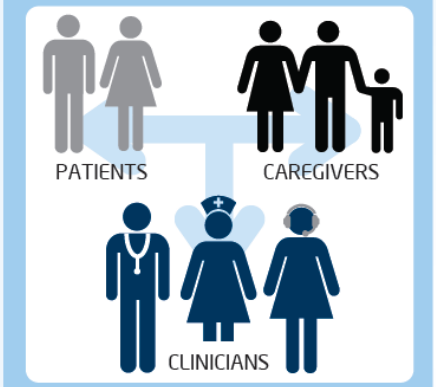
CONNECT

All eyes on the same, shared information

- Electronic health records (EHRs)
- Personal health records (PHRs)
- Security from cell to cloud
- Health information exchange (HIE) software
- Ubiquitous, fast wireless

COORDINATE

Team-based care and collaboration for care and pay



- Online team portals
- Care plan creation and status tools
- Real-time status dashboards
- Quality reporting tools and cycles
- Shared payment and asset tracking

SUPPORT

Decision support from surgeons to citizens

- Algorithms for real-time and recursive information processing
- Clinically validated physician support tools
- Consumer context-aware decision support tools
- Complex, comorbid care management

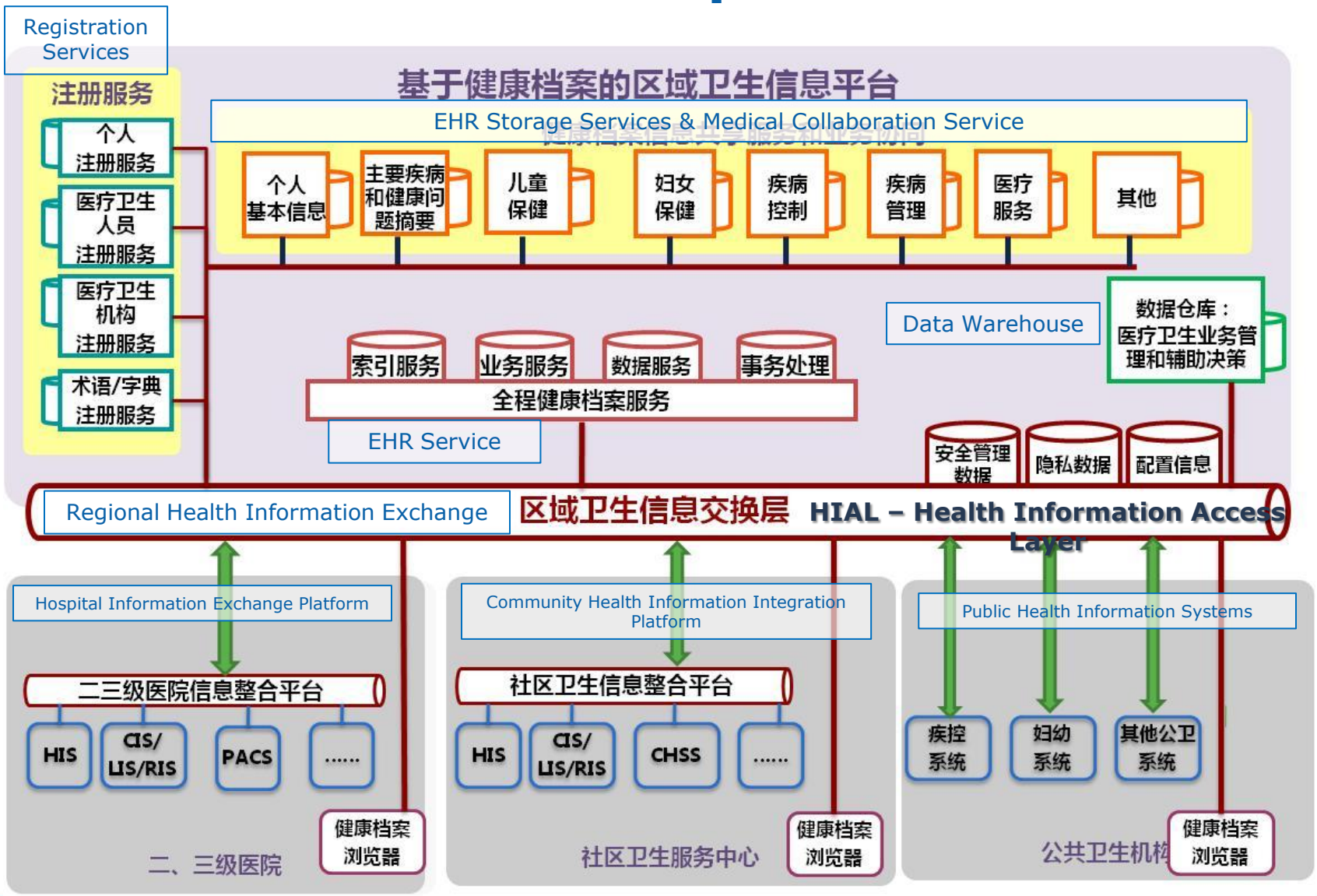
PERSONALIZE

Close the loop with individual, customized care

- Personalized prompting and coaching/mHealth
- Multi-device interusability for care management/CDM
- Real-time feedback on drug and behaviorial therapies
- Reliable real time care intervention

WORKFORCE AND WORKFLOW

China EHR based RHIN platform



Challenges

Various data sources, unstructured, texts, images, videos, etc.

- Health records, lab reports, billing data, PACS images, physical orders, follow-ups, etc

Difficult to standardize the data format

- Data needs to be stored for 50 years, its format keeps changing.
- HL7 Clinical Document Architecture (XML) is evolving frequently.

Big data volume

- 10PB： A medium city in China（10M population）, 50 years' data

Any existing IT system in China cannot process these data in 3~5 years.

Opportunities…

Improving efficiency and reducing costs

- real-time information sharing from clinics, doctors to patients
- real-time status dashboard

Computer aided diagnostics/research

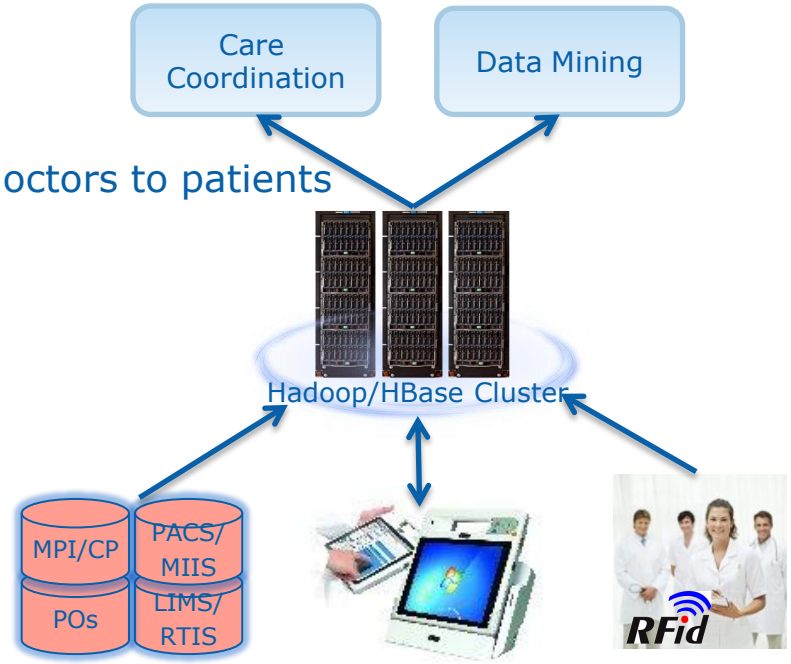
- Disease classifications, like blood poison

Decision support system

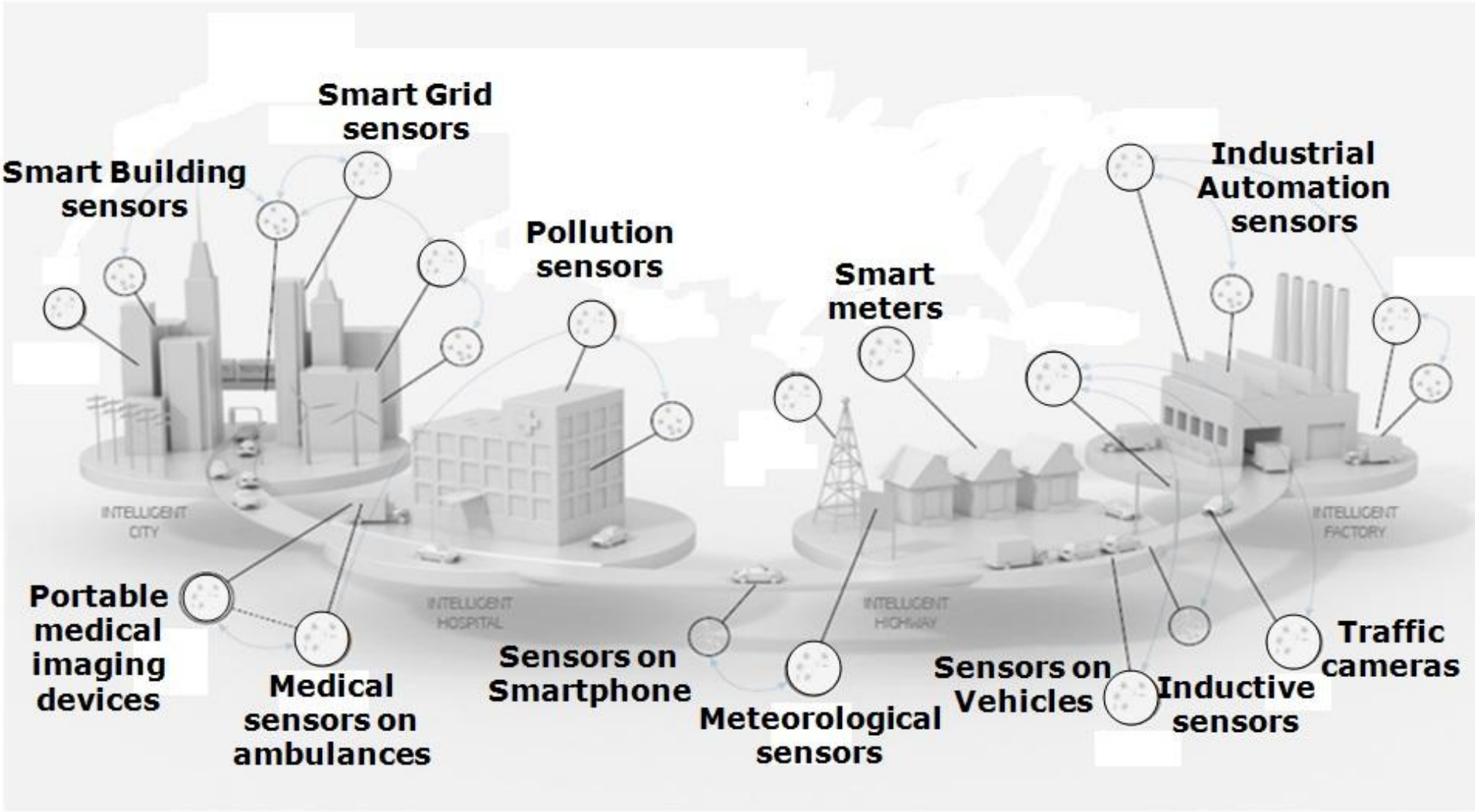
- Trend analysis：cancer trend analysis、epidemic disease analysis、etc
- Association analysis：adverse drug events analysis

Personalized Medicine

- Personalized prompting and coaching



Internet of Things



Challenges

Numerous data

- City A: 500,000 cameras, 200PB video within 3 months
- City B: 12,000 ITS cameras, 2B traffic records per day, 1PB records in 3 months

Real-time processing

- Real-time data collection, scan, query and sharing
- Real-time event detection
- Near real-time predictive analysis

Large scale distributed processing

- Central data center is not affordable, because of money, space, power supply, air conditioner, etc
- Application needs a uniform way to access the data

HBase as the infrastructure, but needs:

Global Table View

- Geographical distributed DCs, connected through high speed network
- One very big table across multiple data centers

Active-Active Availability

- Available for read/write even in case of data center failure(s)

Durability

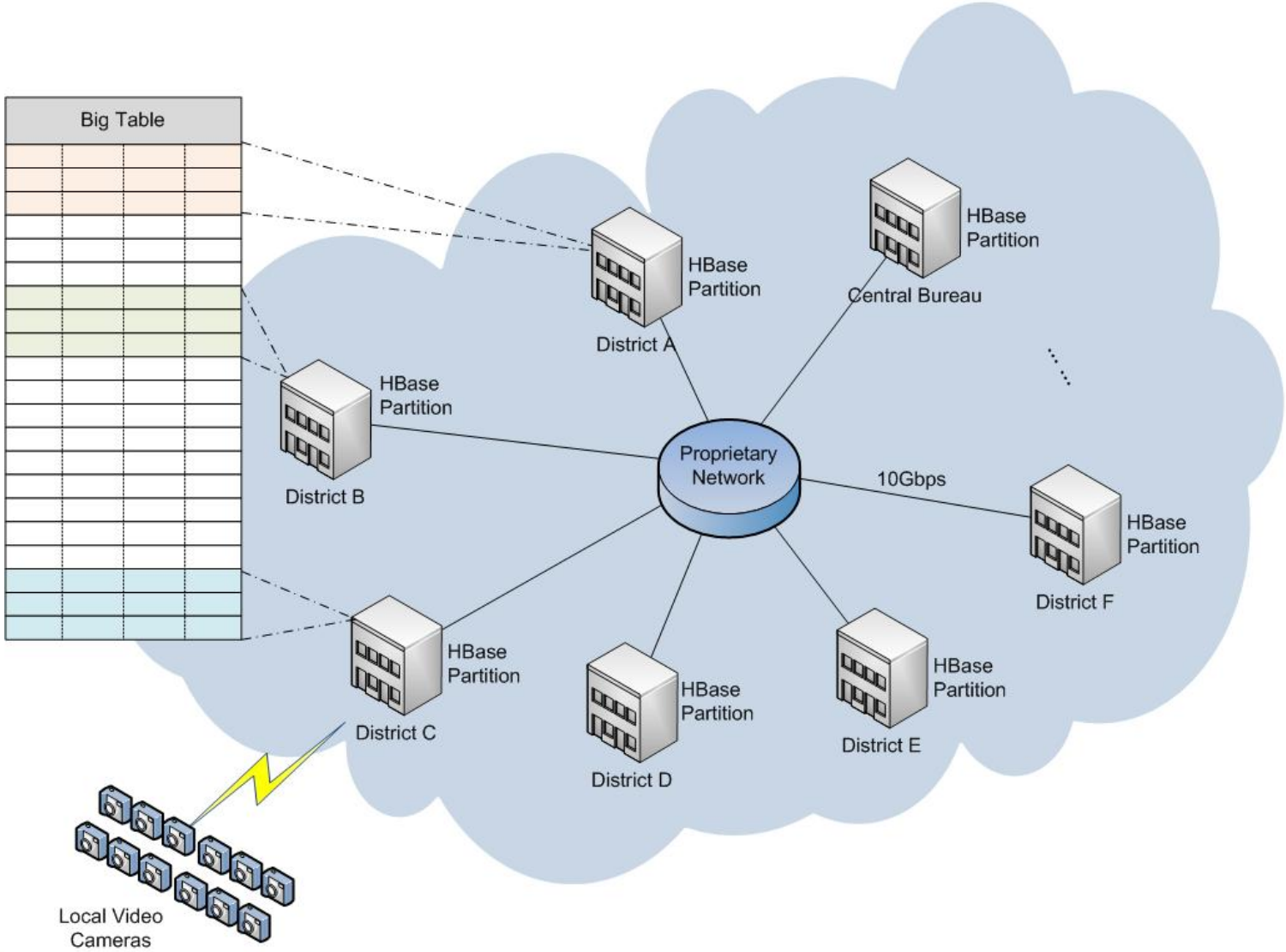
- Auto recover from data center failure

Locality

- Reduce write latency
- Reduce network bandwidth requirement

Eventual Consistency across data centers

Big table over DCs (a reference architecture)




intel
Amazing things happen with Intel inside®