Data Protection and Management

Big Data Data Protection

ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

What is Big Data

- ## Big data is larger
  - Volume in heterogeneous formats
  - Unstructured
  - Growing rapidly
  - Cannot use traditional tools and approaches to analyze and process

- ## More complex data sets

	Traditional Database	Big Data
Flexibility	Fixed schema that is static in nature	Dynamic schema that can include structured as well as unstructured data
	Structured data that fit effortlessly into relational databases or tables	Data is stored in a raw form and the schema is applied only when accessing it.
		Datasets from diverse sources are appended
Real-time analytics	Always took place after the event or time period that was being analyzed	Analytics takes place in real-time — as the data is being gathered — and findings are presented practically instantaneously
Distributed architecture	Based on a centralized database architecture	Bses a distributed architecture
		Computation is distributed among several computers in a network
Multitude of sources	The sources of data were fairly limited	Multitude of sources that capture data practically every moment
Enables exploratory analysis	Need to determine their questions at the start	Enables a more iterative and exploratory approach
	Led to the creation of monthly reports, productivity analysis, customer survey findings	Provides insights into sentiment analysis, product strategy, asset utilization, preventive maintenance of equipment, etc

Characteristics of Big Data

Volume
- Massive volumes of data
- Challenges in storage and analysis

Velocity
- Rapidly changing data
- Challenges in real-time analysis

Variety
- Diverse data from numerous sources
- Challenges in integration and analysis

Variability
- Constantly changing meaning of data
- Challenges in gathering and interpretation

Veracity
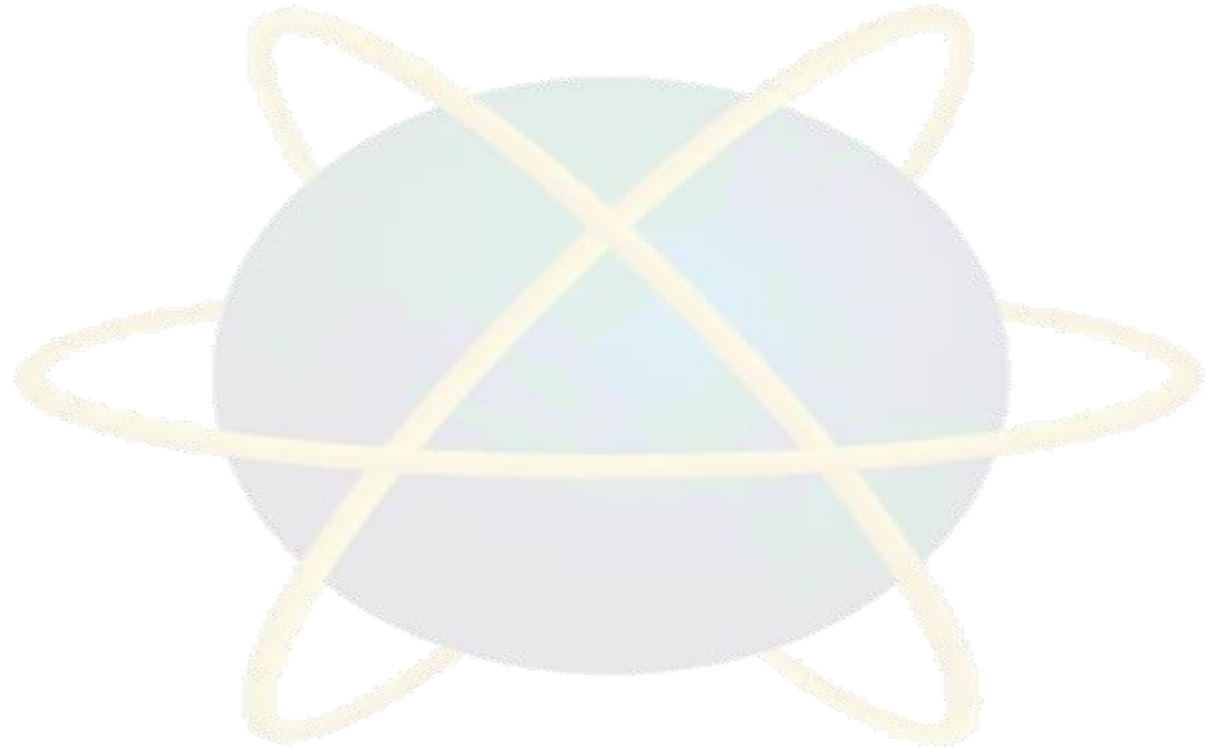- Reliability and accuracy of data
- Challenges in transforming and trusting data

Value
- Cost-effectiveness and business value

Quick Review

- What should be done by expert in order to reap all the benefits of big data analytics

Why Big Data Analytics?

Business Driver	Examples
Desire to optimize business operations	Sales, pricing, profitability, efficiency
Desire to identify business risk	Loss of customer, fraud, default
Predict promising new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-money laundering, Fair lending, Basel II-III, Sarbanes-Oxley (SOX)

Components of a Big Data Analytics Solution

Query
- Efficient way to process, store and retrieve data
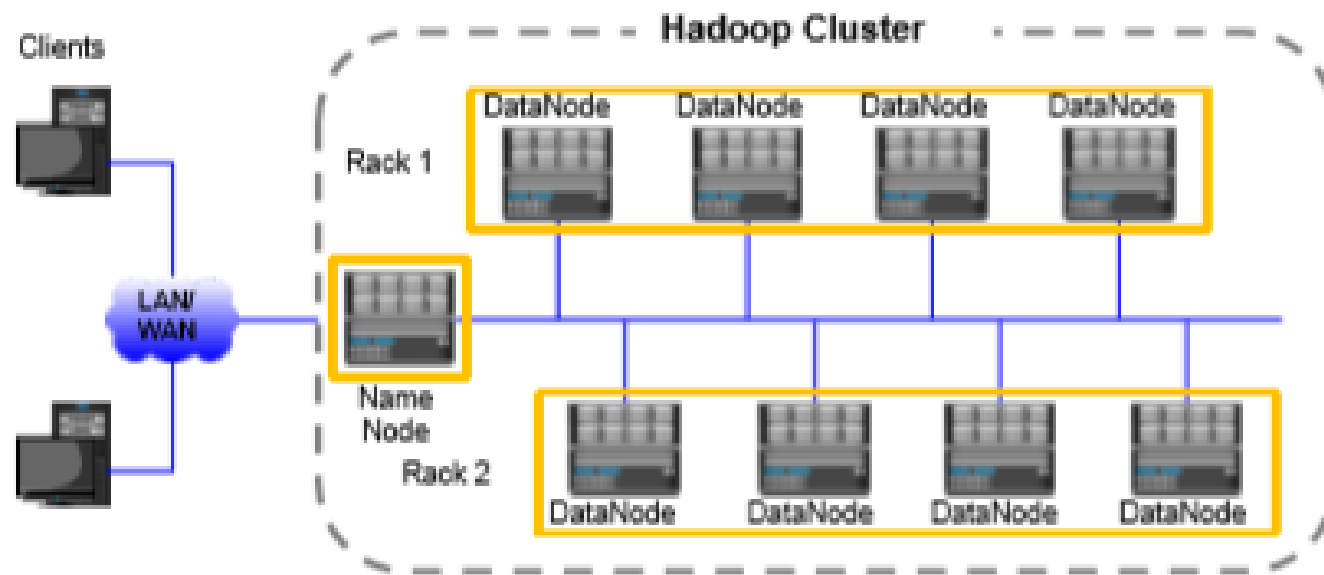- Platform for user-friendly analytics systems

MapReduce
- Distributes (parallel) computation over many servers
- Batch processing model

Storage
- Distributed architecture (HDFS)
- Non-relational, Unstructured content

Hadoop Distributed File System (HDFS)

HDFS architecture has two key components

- Name Node
  - Acts as a primary server and has in-memory maps of every file, file locations, as well as all the blocks within the file and the DataNodes to which they reside on
  - Responsible for managing FS namespace and controlling the access of files by the clients

- Data Node
  - Acts as secondary that serves R/W requests as well as performs block creation, deletion, and replication

Big Data Protection Challenges

- Need to protect massive volumes of data which exceeds the capabilities of traditional data protection solutions

- Hard to find out what data needs to be protected

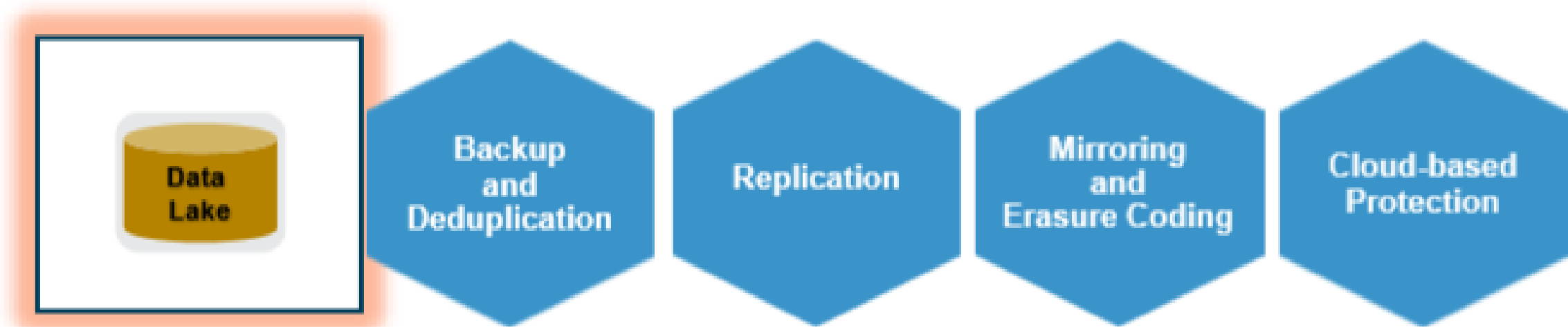- More data may affect the service level agreements

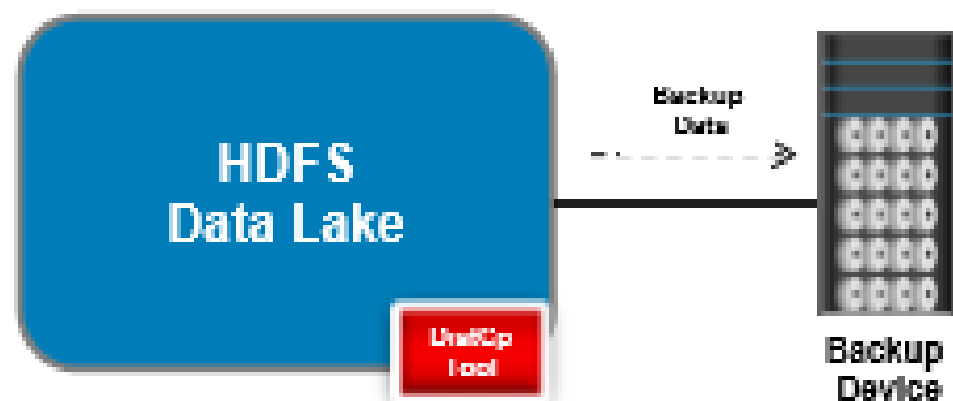- Requires seamless integration of data repository (data lake) with data protection software

- Difficult to protect the data within the budget

Data Lake
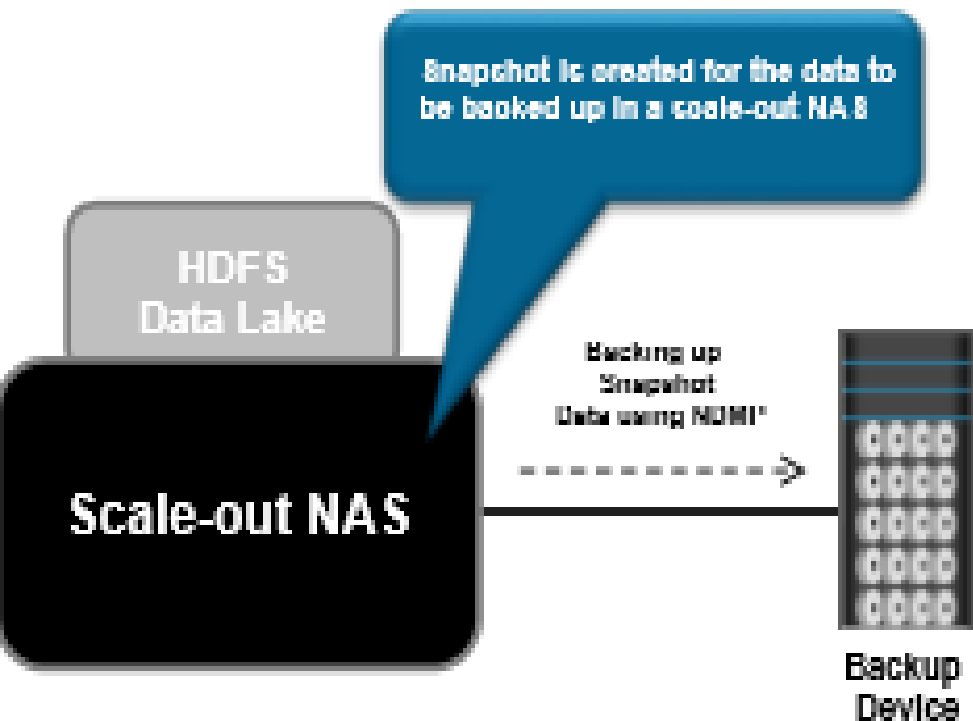
Data Lake Protection – Backup

Hadoop Distributed Copy (DistCp)

- DistCp is a native utility, built into HDFS to backup and restore data from data lake to a backup device

- Two approaches for backing up data
    1. Takes a file system snapshot and performs backup using DistCp tool from the snapshot
    2. Takes backup directly from the HDFS data lake to the backup device using DistCp tool

Data Lake Protection – Backup

Backup of Snapshot created in a Scale-out NAS using NDMP

Snapshot is created for the data to be backed up in a scale-out NA S

HDFS Data Lake

Scale-out NAS

Backing up Snapshot Data using NDMP

- - - - - - - - - - - ->

Backup Device

- HDFS data lake is created in a scale-out NAS

- Snapshot is created for the data to be protected and is backed up using NDMP

- Snapshot data can be backed up to a backup device – scale-out NAS, and scale-out object storage

Big Data as a Service

- Service provider offers resources to enable the consumers to run big data analytics workload in the cloud
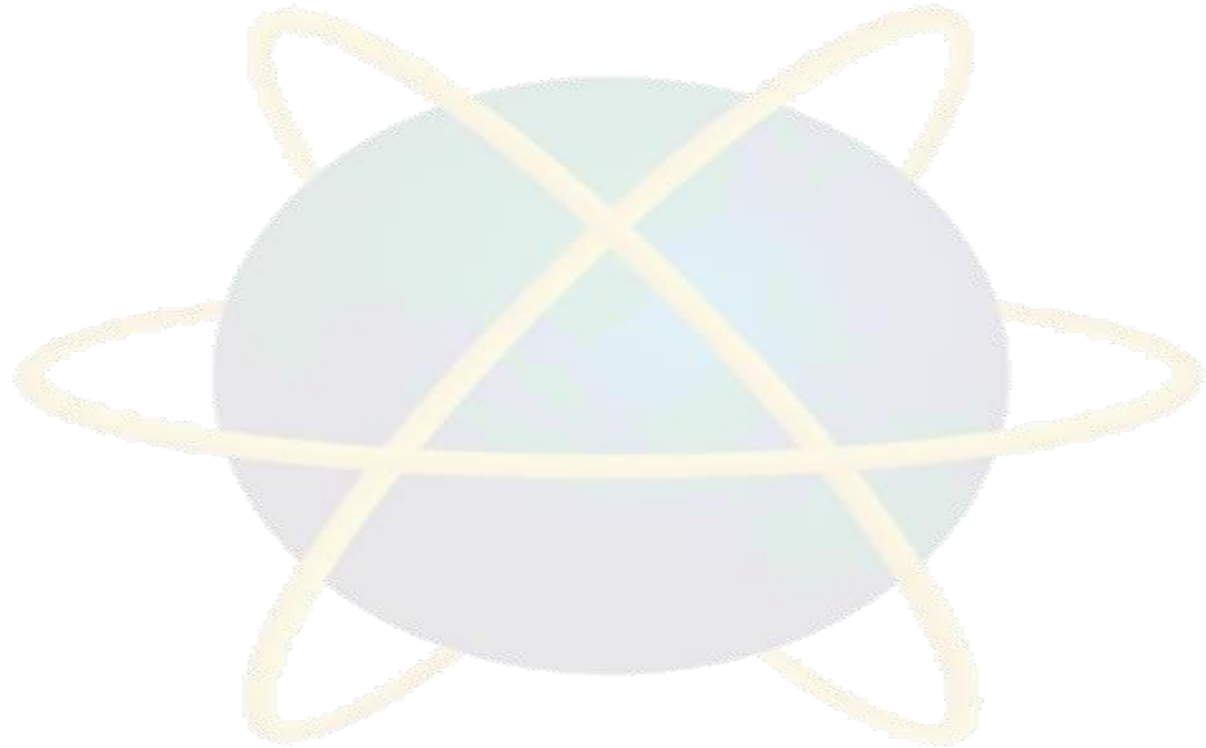

Big Data as a Service Models

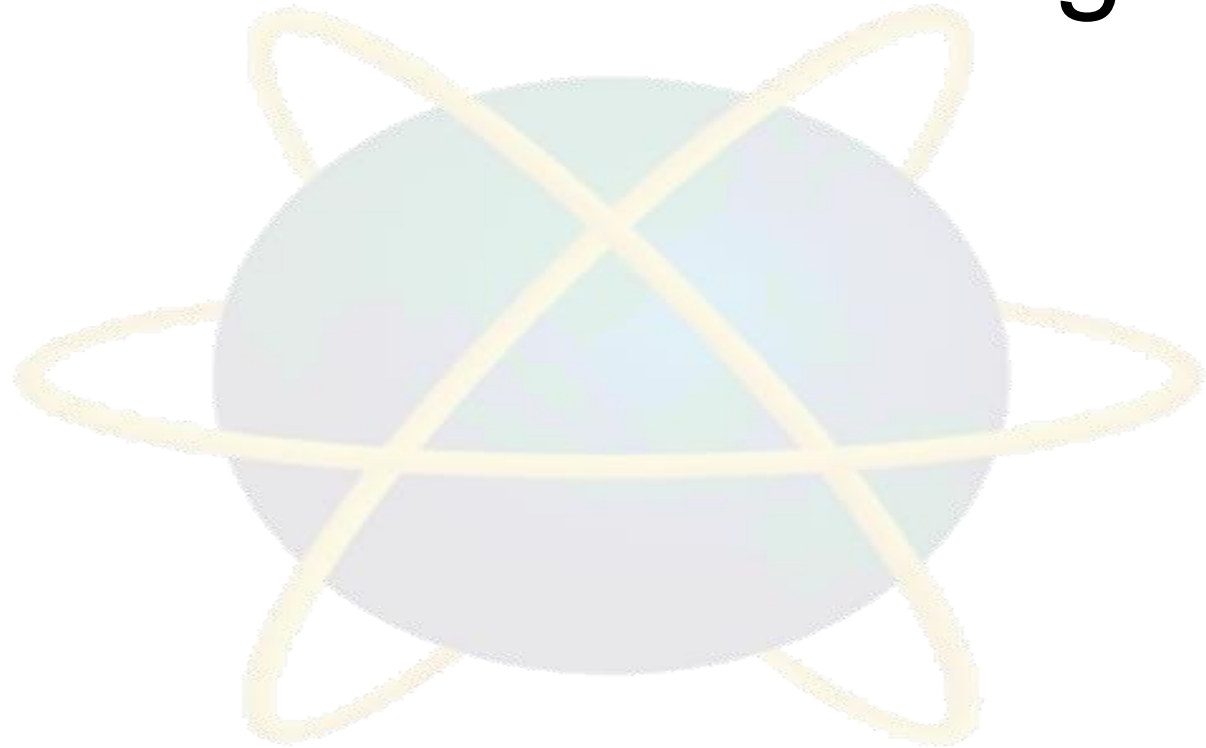Big Data Infrastructure as a Service

Big Data Platform as a Service

Big Data Analytics Software as a Service

Quick Review

- Do a google search to identify and explain the essential big data practices for business.
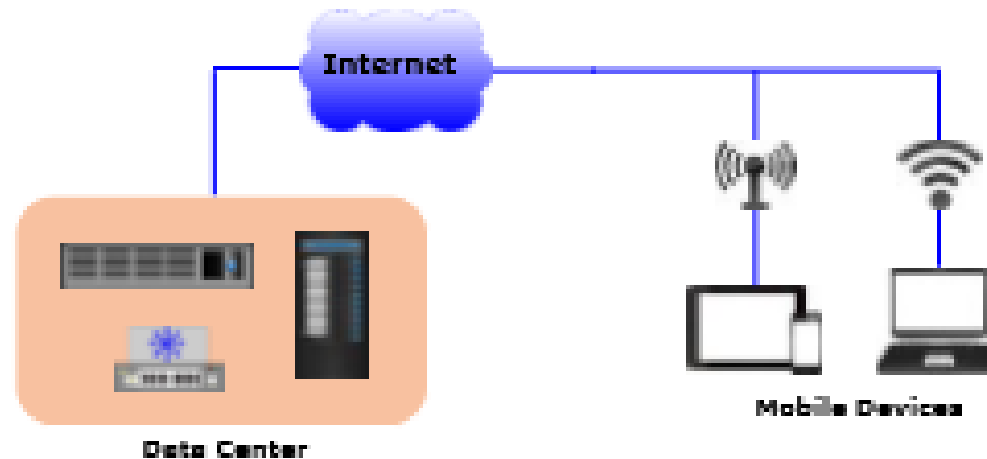
Protecting Mobile Devices

Mobile Device Overview

A compute system that is portable and typically a handheld device with a display, and has either a keyboard and/or touch input

Examples of Mobile Device
- ✓ Laptop
- ✓ Tablet
- ✓ Smartphones
- ✓ Personal Digital Assistants

- Enables users to access applications and information from their personal devices from any location

- Increases collaboration and enhances workforce productivity

Key Challenges in Protecting Mobile Device Data

- Data is protected (backed up) only when the mobile device is online

- Data protection from mobile device to data center is impacted due to intermittent network connectivity

- Devices are not always connected to the corporate network, so it connects over the Internet, which may rise to security threat

- Data protection software must support mobile device OS

- Network bandwidth limitation

- Security features on the mobile devices restrict the access of the data stored on the device

To overcome these challenges, organizations must adopt new policies, strategies, and techniques to protect the data residing on mobile devices

Mobile Device Backup

Mobile Backup Clients

Backup Data

WAN

Enterprise Data Center
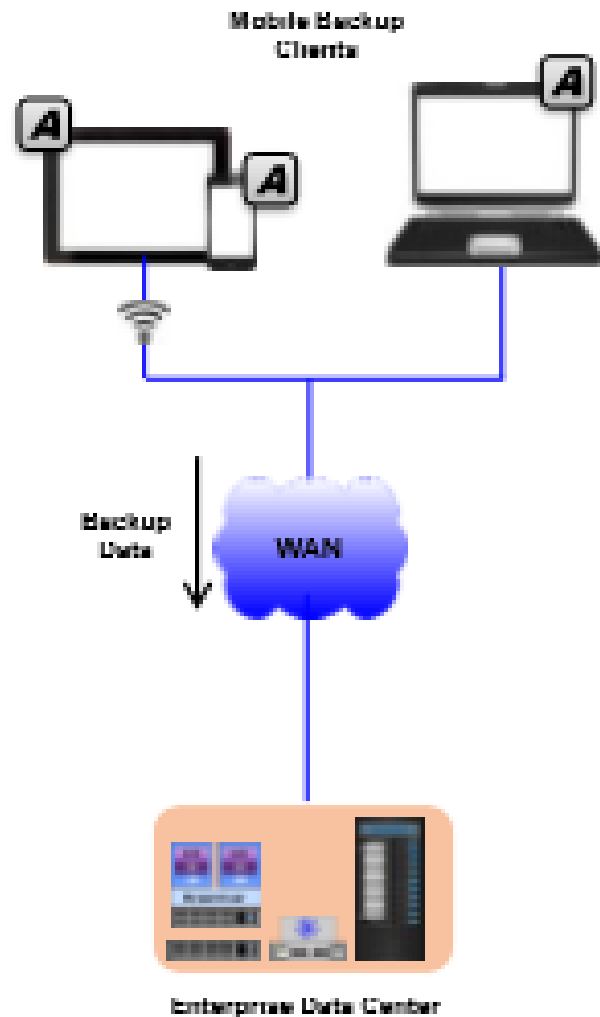
- Requires installing backup client application (agent) on the mobile devices
  - Backs up the data to the enterprise data center

- Data can be backed up manually or automatically from mobile devices

- Deduplication, compression, encryption, and incremental backup can be implemented for performing mobile device backup
  - Provides network and backup storage optimization, and security

File Sync-and-Share Application
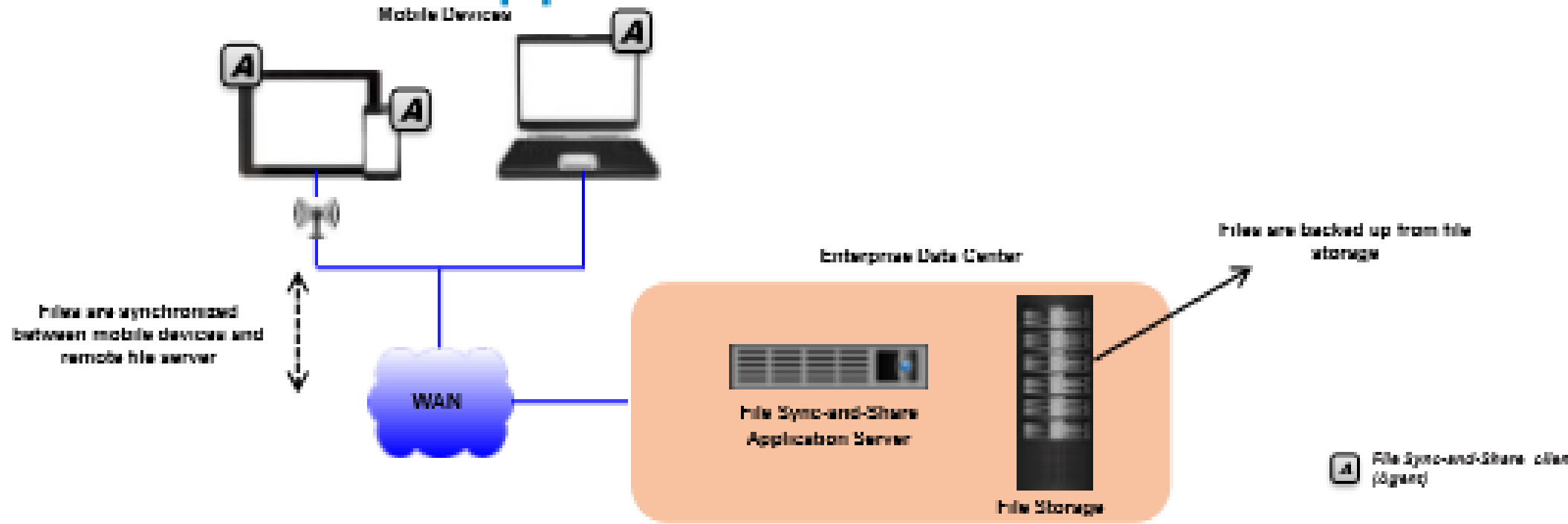
- Automatically establishes two-way synchronization between the device and a designated network location (enterprise data center)

- Files are backed up from the remote storage instead of the mobile devices

- Improves productivity by allowing users to access data from any device, anywhere, at any time

Mobile Cloud Computing
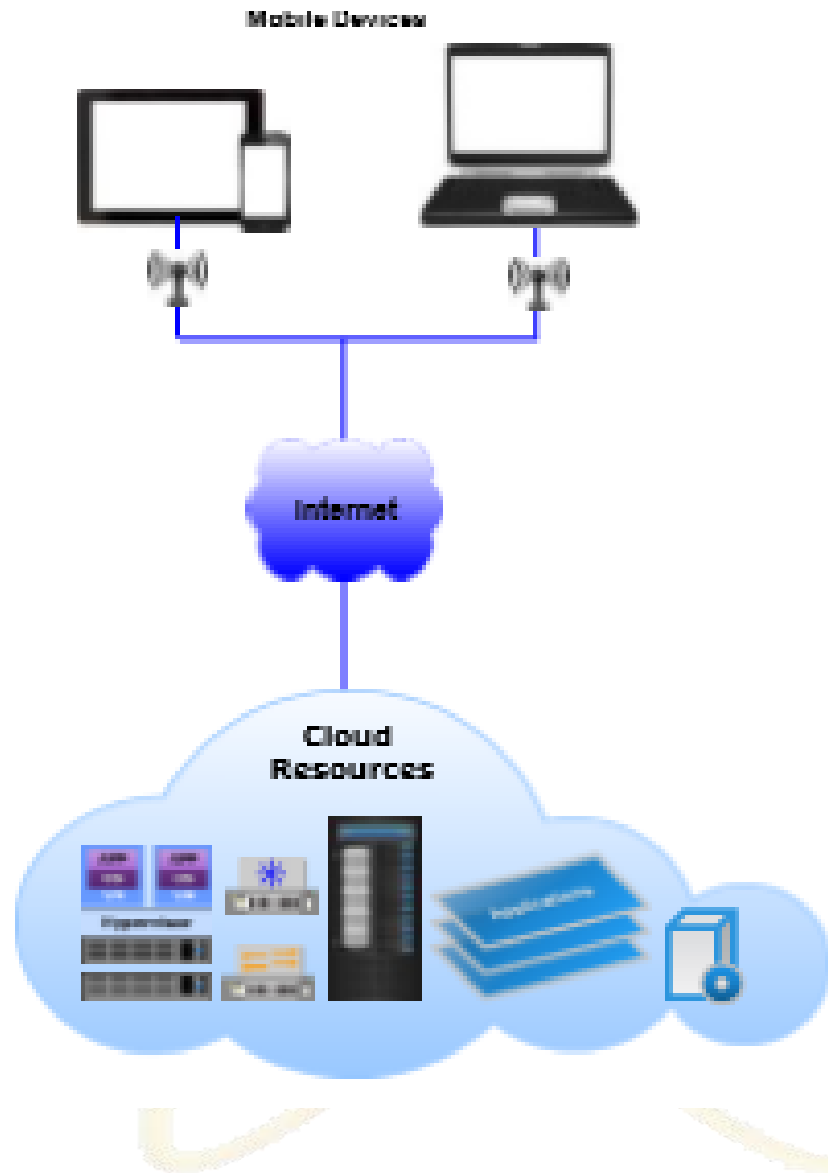
- Compute processing and storage are moved away from the mobile device and takes place in a computing platform located in the cloud

- Applications running in the cloud are accessed over wireless connection using a thin client application/web browser on the mobile devices

- Cloud services accessed over mobile devices
 - SaaS examples: Cloud storage, travel and expense management, and CRM

Cloud-based Mobile Device Data Protection

Cloud-based Backup
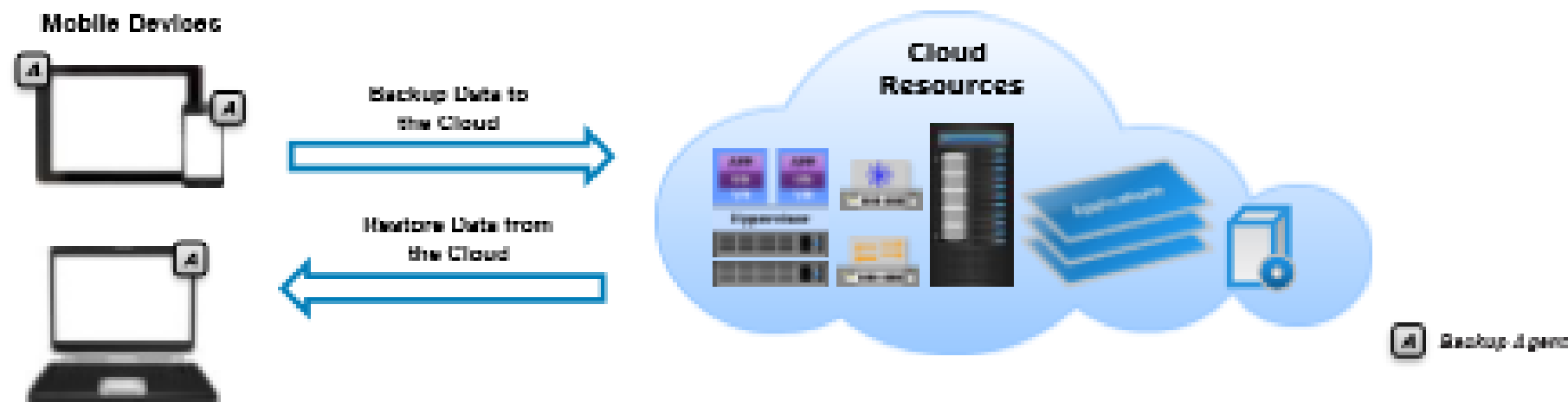
- Backup client application (agent) that is installed on the device enables to access and perform backup to the cloud
  - Typically backs up only the changed blocks to the cloud storage

- Some mobile applications have built-in backup feature that backs up the data to the cloud

- Most of the cloud backup solutions available today offer a self-service   portal that allows users to recover data without manual intervention

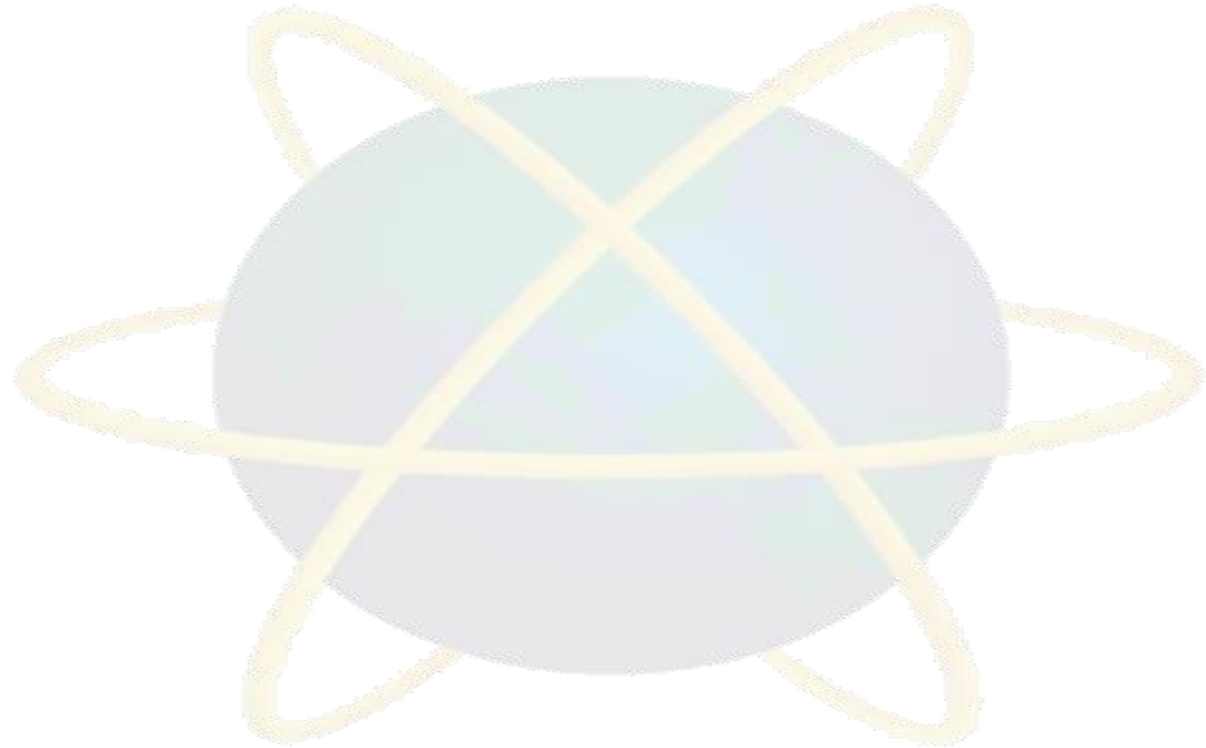- Benefits of Cloud-based Backup for Mobile Devices?
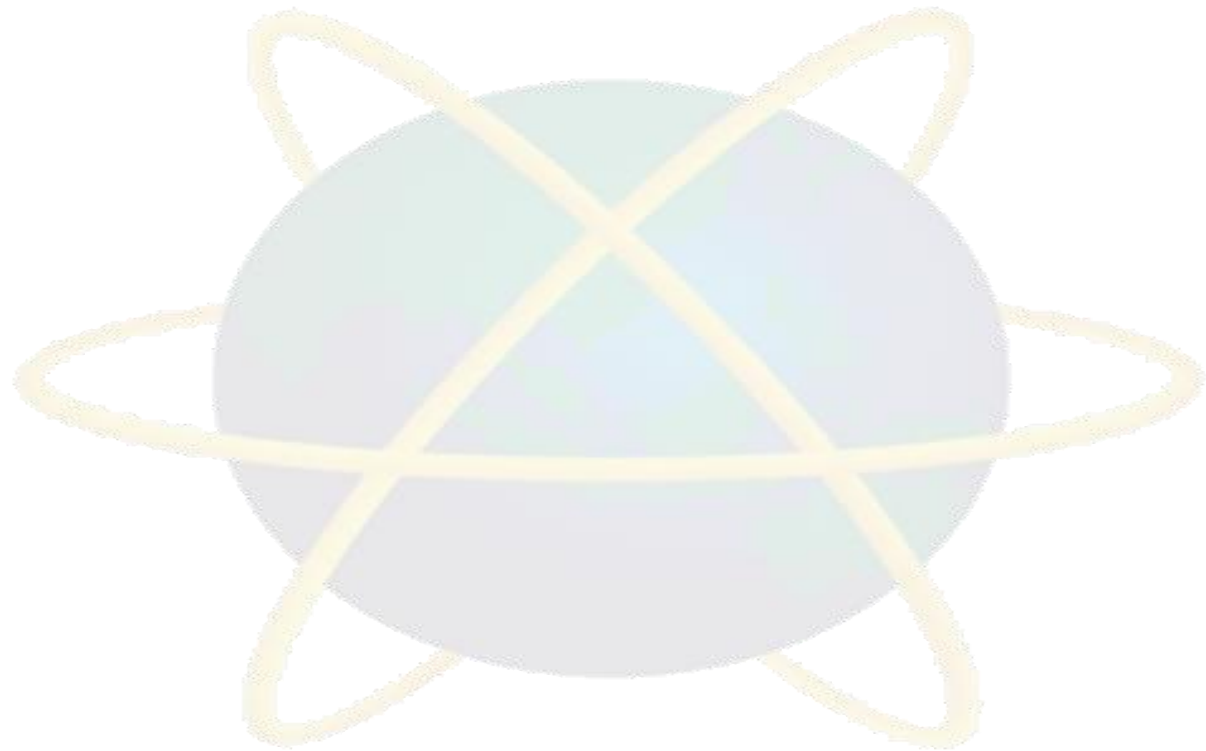
Concepts in Practice

- Search for:
- Pivotal GemFire
  - Pivotal Greenplum

visit www.pivotal.io

https://www.youtube.com/watch?v=1sLCxd6CnlA

Q & A

Q & A