Analytics for Object Storage Simplified - Unified File and Object for Hadoop

Sandeep R Patil
STSM, Master Inventor, IBM Spectrum Scale

Smita Raut
Object Development Lead, IBM Spectrum Scale

Acknowledgement : Bill Owen, Tomer Perry, Dean Hildebrand, Piyush Chaudhary, Yong Zeng, Wei Gong, Theodore Hoover Jr, Muthuannamalai Muthiah.

Agenda

- Part 1 : Need as well as Design Points for Unified File and Object
  - Introduction to Object Storage
  - Unified File & Object Access
  - Use Cases Enabled By UFO
- Part 2: Analytics with Unified File and Object
  - Big Data Analytics and Challanges
  - Design Points, Approach and Solution
  - Unified File & Object Store and Congnitive Computing

IBM

Part 1 : Need as well as Design Points for Unified File and Object

- Object Storage Introduction

Introduction to Object Store

- Object storage is highly available, distributed, eventually consistent storage.
- Data is stored as individual objects with unique identifier
- Flat addressing scheme that allows for greater scalability
- Has simpler data management and access
    - REST-based data access
    - Simple atomic operations:
        - PUT, POST, GET, DELETE
- Usually software based that runs on commodity hardware
- Capable of scaling to 100s of petabytes
- Uses replication and/or erasure coding for availability instead of RAID
- Access over RESTful API over HTTP, which is a great fit for cloud and mobile applications
    - Amazon S3, Swift, CDMI API

Object Storage Enables The Next Generation of Data Management

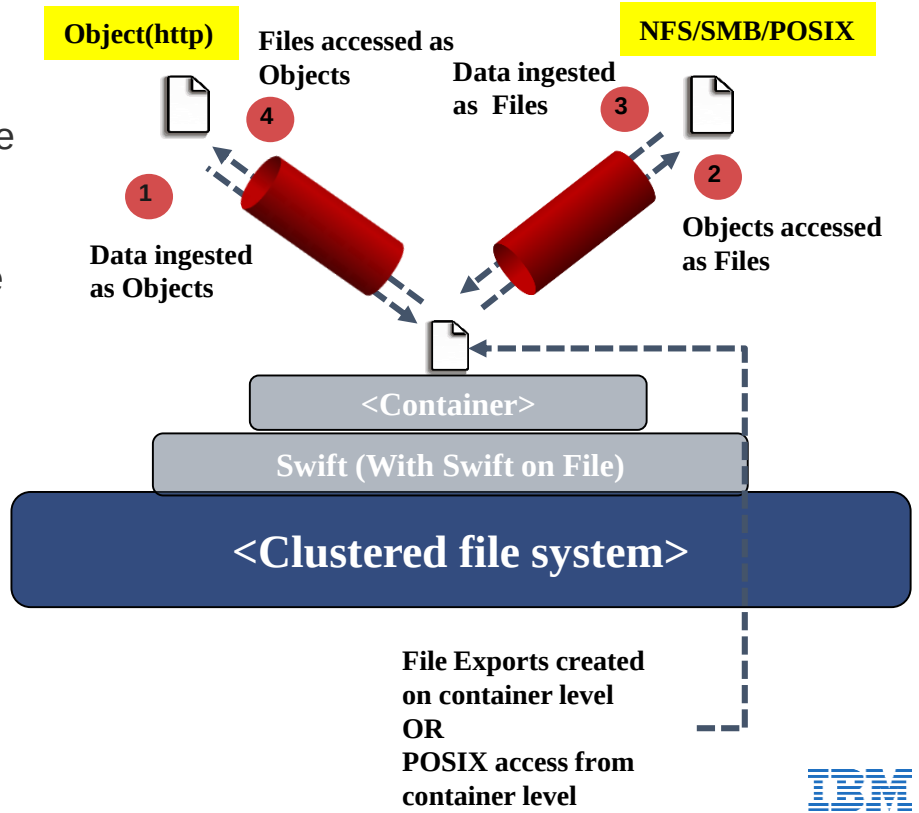
IBM

But Does it Create Yet Another Storage Island in Your Data Center…??

➢ Unified File and Object Access

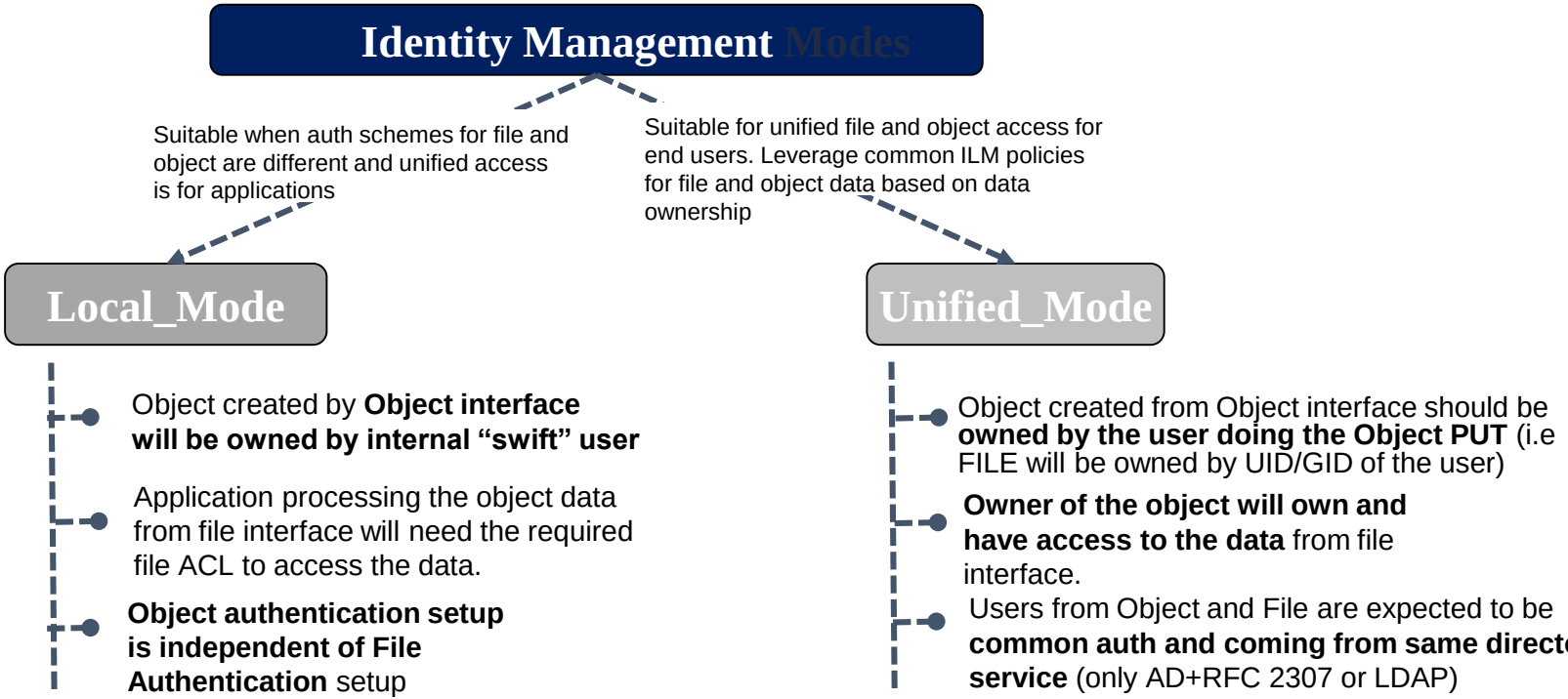
What is Unified File and Object Access ?

- Accessing object using file interfaces (SMB/NFS/POSIX) and accessing file using object interfaces (REST) **helps legacy applications** designed for file to seamlessly start integrating into the object world.
- It allows **object data to be accessed using applications designed to process files**. It allows file data to be published as objects.
- Multi protocol access for file and object in **the same namespace** (with common User ID management capability) allows supporting and hosting data oceans of different types of data with multiple access options.
- Optimizes various use cases and solution architectures resulting in better efficiency as well as cost savings.



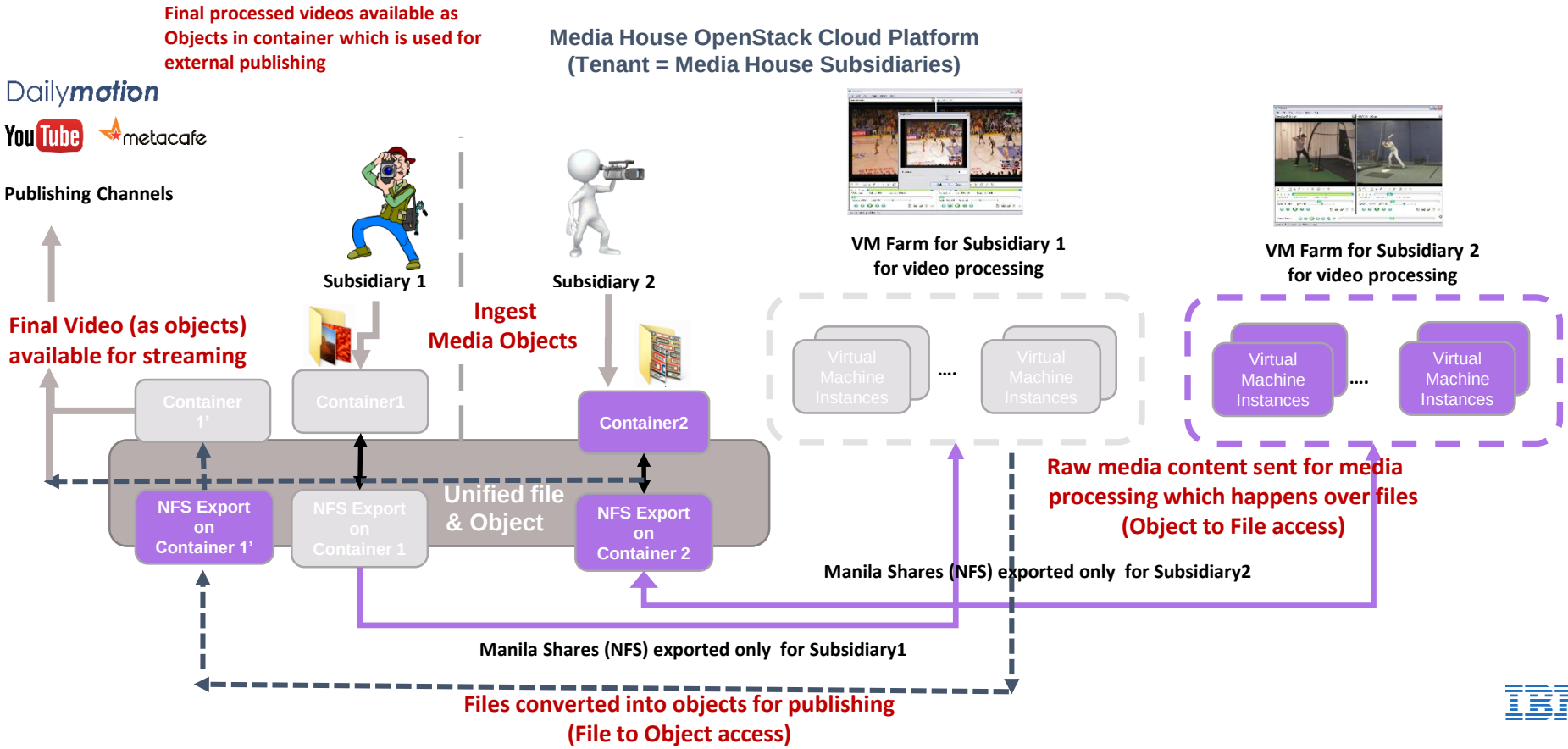
Flexible Identity Management Modes

- Two Identity Management Modes
- Administrators can choose based on their need and use-case



➢ Use Case Enabled by Unified File Object

Use case : Process Object Data with File-Oriented Applications and Publish Outcomes as Objects



We have now understood

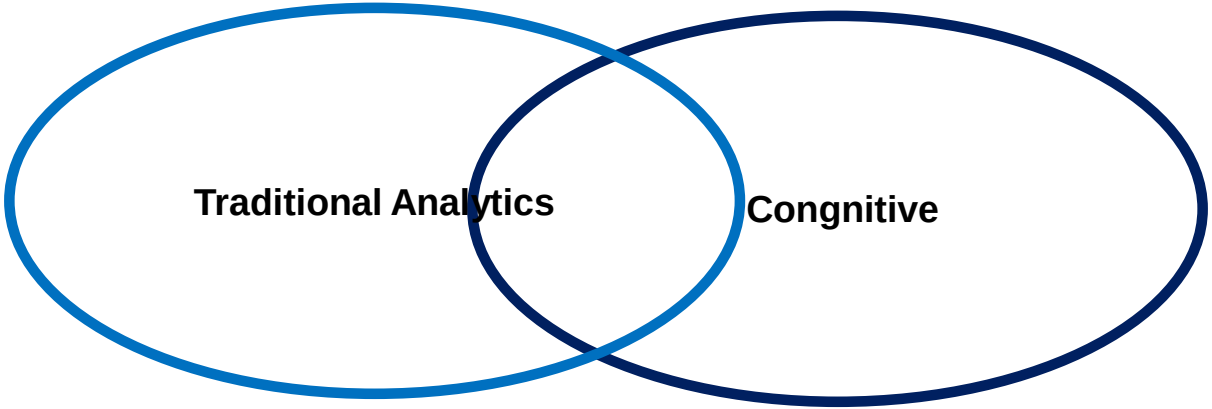
Part 1: Need as well as Design for Unified File and Object

….. Let us now deep dive on

Part 2: Analytics with Unified File and Object

➢ Big Data Analytics and Challenges

Analytics – Broadly Categorized Into Two Sets

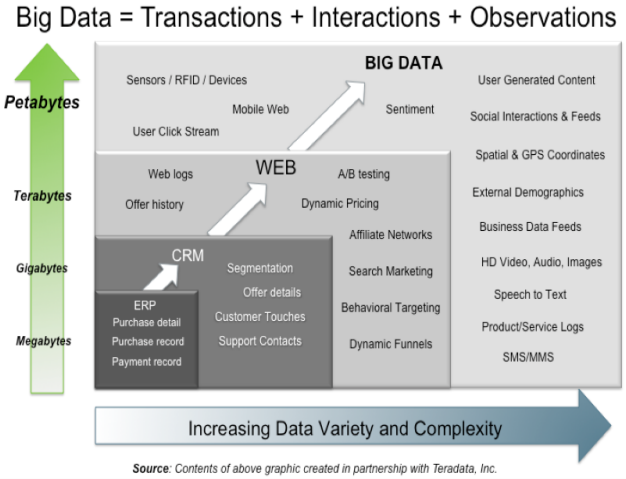


Big Data

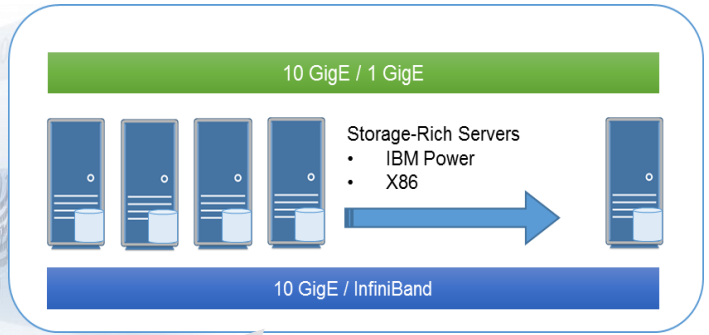
- Big data is a term for data sets that are so large or complex that traditional data processing applications (database management tools or traditional data processing applications) are inadequate.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

Characteristics

- Volume
  - The quantity of generated and stored data.
- Variety
  - The type and nature of the data.
- Velocity
  - Speed at which the data is generated and processed
- Variability
  - Inconsistency of data sets can hamper processes manage it
- Veracity
  - Quality of captured data can vary greatly, affecting accuracy



Challenges with the Early Big Data Storage Models



Key Business processes are now depend on the analytics

It's not just one type of analytics

hadoop

Spark

sas

SAP HANA

Ingest data at various end points

More data source than ever before, not just data you own, but public or rented data

Move data to the analytics engine

It takes hours or days to move the data!

Perform analytics

Can't just throw away data due to regulations or business requirement

Repeat!

IBM

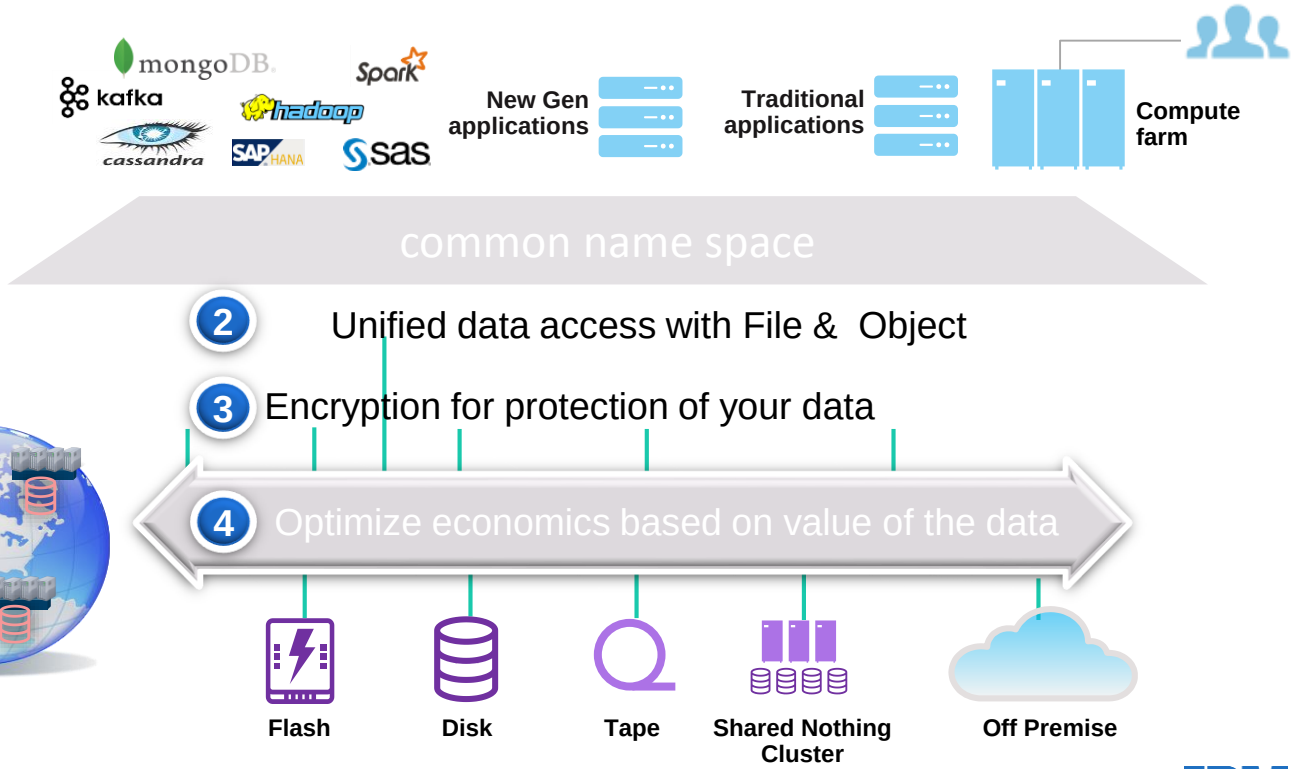
➢ Design Points, Approach & Solution

What are the Solution Design Points that we came across?

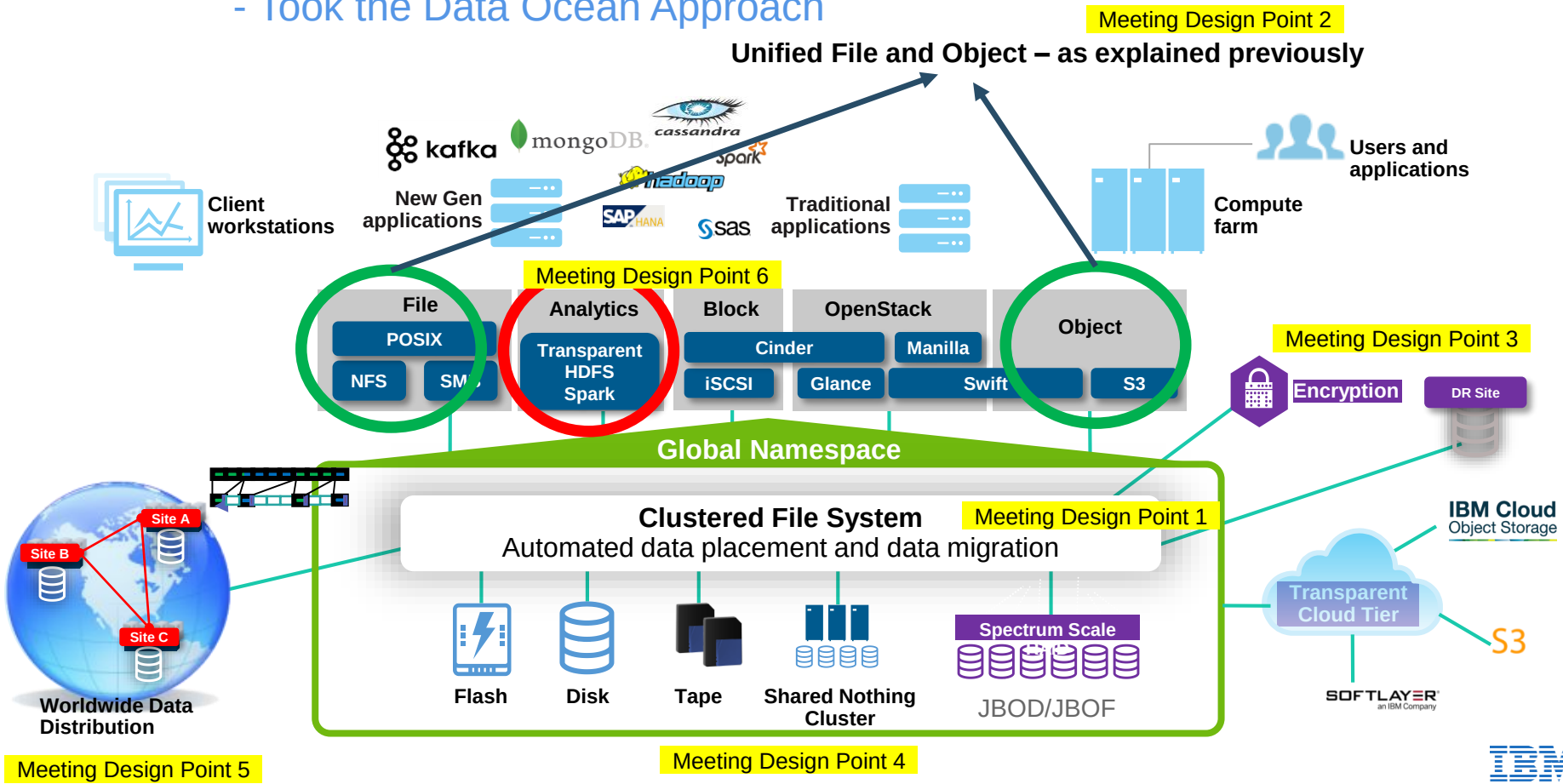
6 Bring analytics to the data

1 Single Name Space to house all your data (Files and Object)

5 Geographically dispersed management of data including disaster recovery



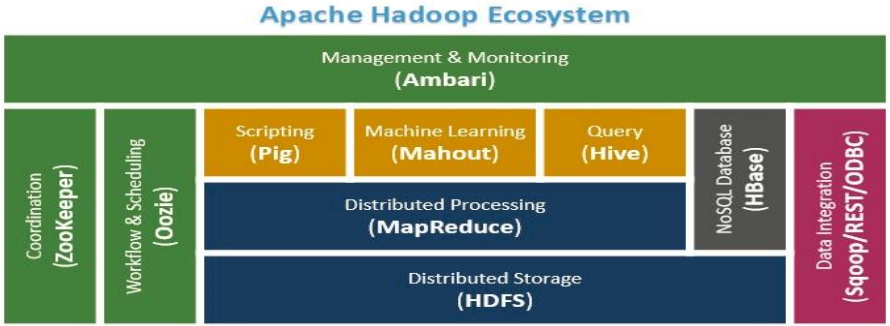
How Did We Approach The Solution & address the Design Points?
- Took the Data Ocean Approach



Meeting Design Point 6 – Bring Analytics to Data

Apache Hadoop - Key Platform for Big Data and Analytics

- An open-source software framework and most popular BD&A platform
- Designed for distributed storage and processing of very large data sets on computer clusters built from commodity hardware
- Core of Hadoop consists of
  - A processing part called MapReduce
  - A storage part, known as Hadoop Distributed File System (HDFS)
  - Hadoop common libraries and components
- Leading Hadoop Distro: HortonWorks, CloudEra, MapR, IBM IOP/BigInsights



Meeting Design Point 6 – Bring Analytics to Data
HDFS Shortcomings

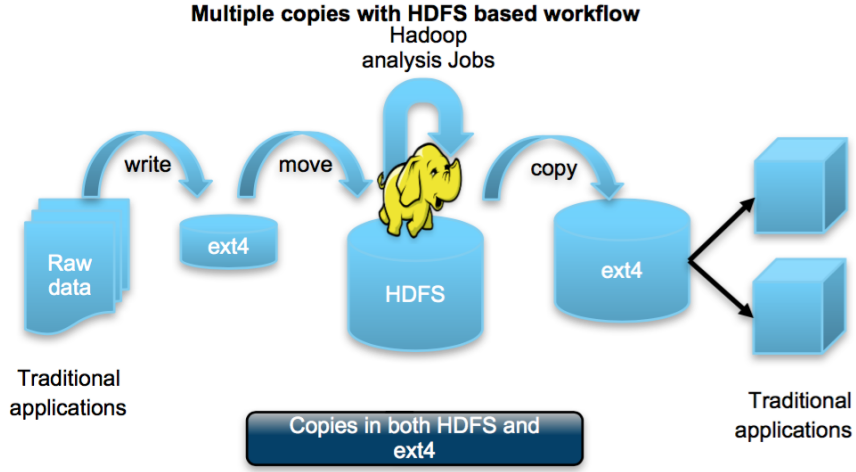
- HDFS is a **shared nothing architecture**, which is **very inefficient for high throughput jobs** (disks and cores grow in same ratio)
- **Costly data protection**:
  - uses 3-way replication; limited RAID/erasure coding
- Works only with Hadoop i.e **weak support for File or Object protocols**
- Clients **have to copy data from enterprise storage to HDFS** in order to run Hadoop jobs, this can result in running on stale data.

IBM

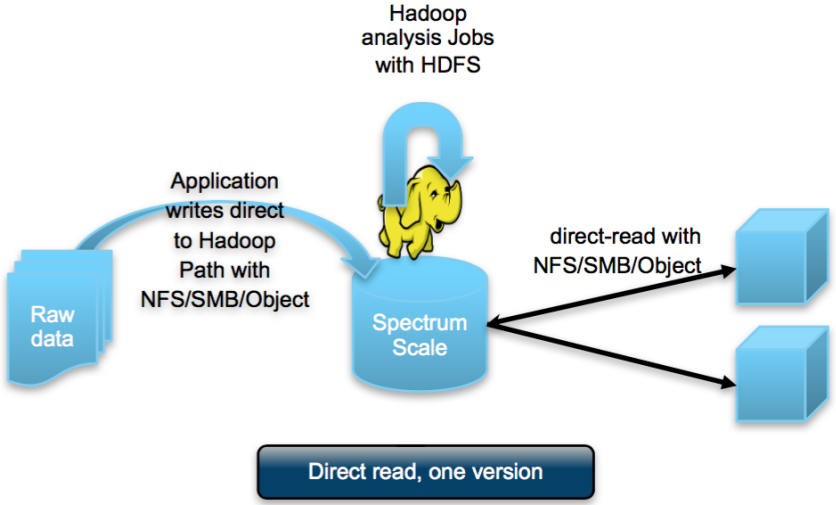
Meeting Design Point 6 – How to Bring Analytics to Data ?

Reduce the datacenter footprint

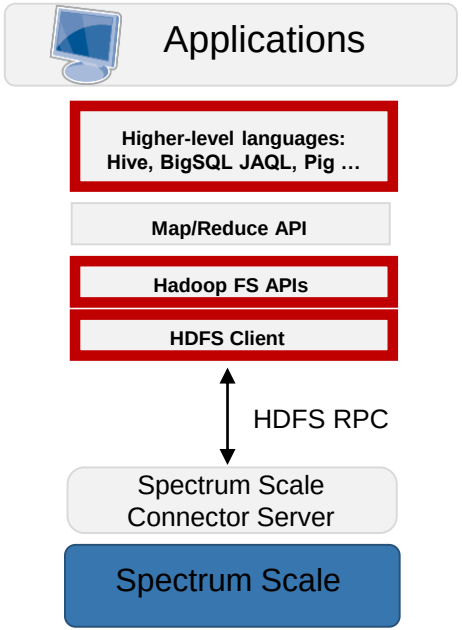
PROBLEM: Data Scientists waste days just copying data to HDFS

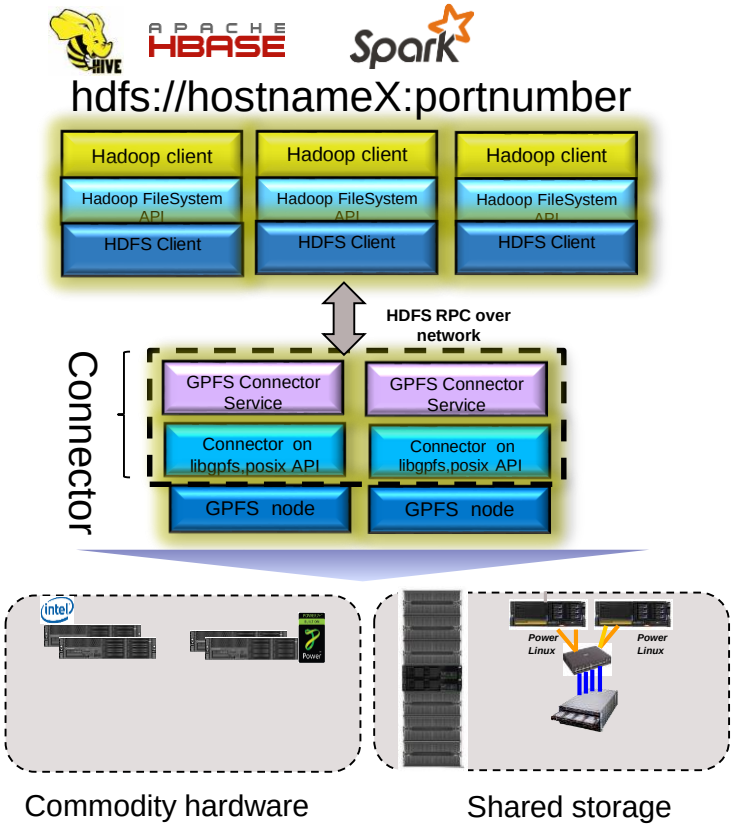


Desired Solution: Need In place Analytics (No Copies Required). Clustered Filesystem should support HDFS Connectors



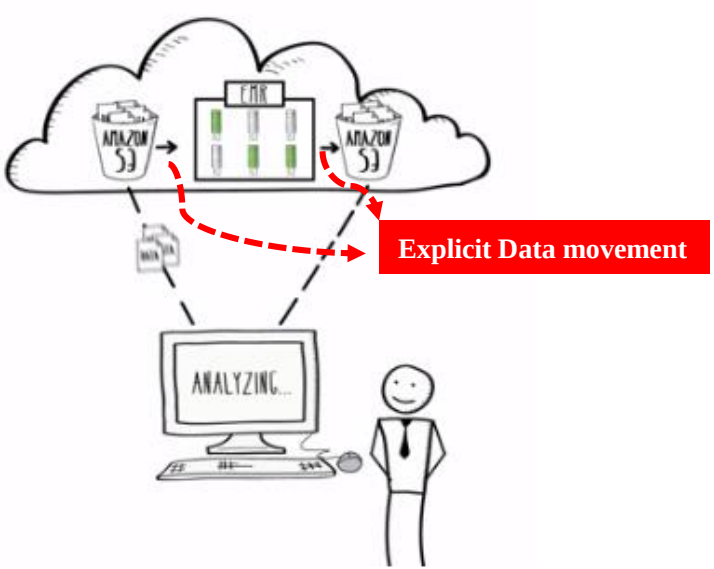
How did we design to overcome the inhibitors: Developing a HDFS Transparency Connector with Unified File and Object Access





Meeting Design Point 6- "In-Place" Analytics for Unified File and Object Data Achieved.

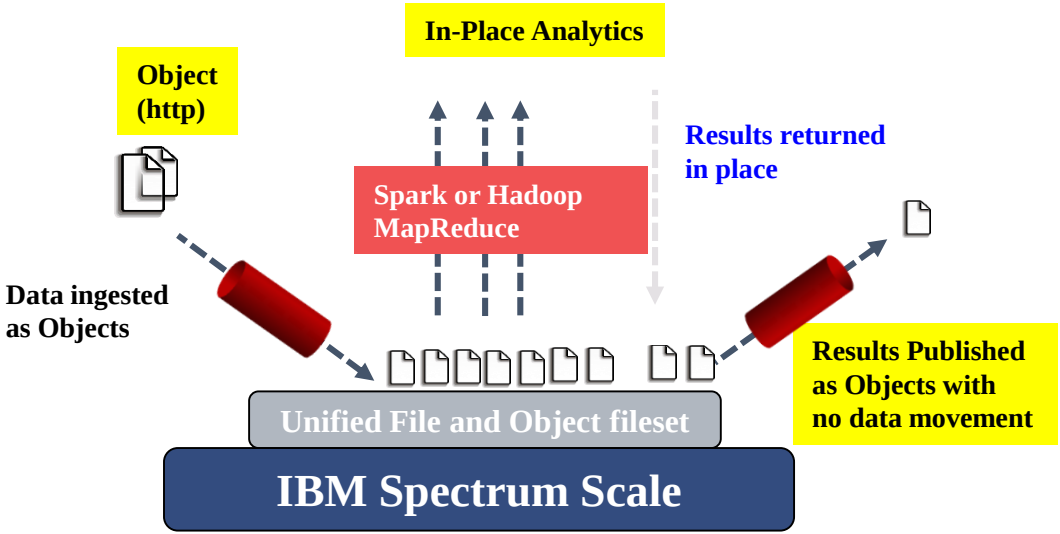
Analytics on Traditional Object Store



Traditional object store – Data to be copied from object store to dedicated cluster , do the analysis and copy the result back to object store for publishing

Source:https://aws.amazon.com/elasticmapreduce/

Analytics With Unified File and Object Access



Object store with Unified File and Object Access –
Object Data available as File on the same fileset. Analytics systems like Hadoop MapReduce or Spark allow the data to be directly leveraged for analytics.

No data movement i.e. In-Place immediate data analytics.

- Unified File & Object Store and Cognitive Computing

Cognitive Services

- Cognitive services are based on a technology that encompasses machine learning, reasoning, natural language processing, speech and vision and more
- IBM Watson Developer Cloud enables cognitive computing features in your app using IBM Watson's Language, Vision, Speech and Data APIs.

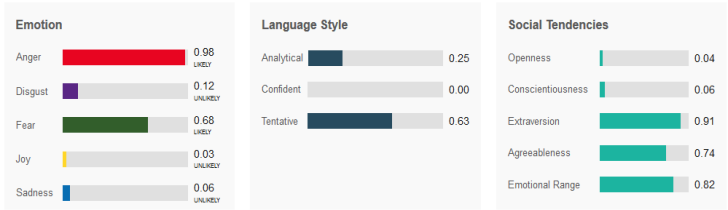
Examples of Watson Services

Visual Recognition

Speech to Text

Input Audio / Voice

Tone Analyzer



Some more…

- Alchemy Language – Text Analysis to give Sentiments of the Document
- Language Translation – Translate and publish content in multiple languages
- Personality insights – Uncover a deeper understanding of people's personality
- Retrieve and Rank – Enhance information retrieval with machine learning
- Natural language Classifier – Interpret and classify natural language with confidence

And Many More …

Cognitive Services help derive the meaning of data

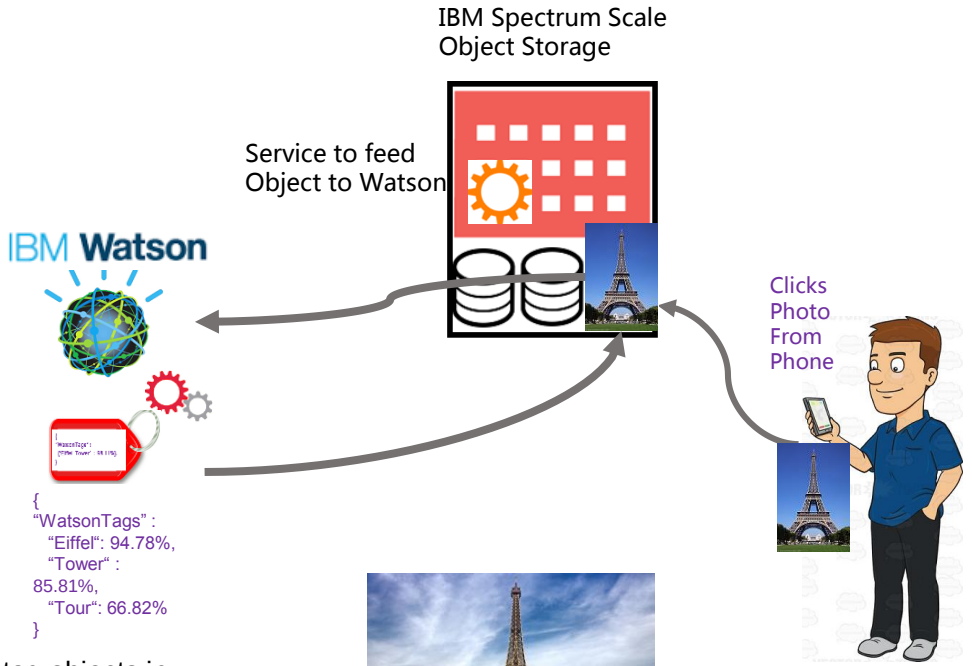
Cognitive Service with Unified File and Object

We need a provision to Cognitively Auto-tag Heterogeneous Unstructured data in form of Object to Leverage its benefits….

SwiftonFile based object storage + IBM® Watson™ Services for Bluemix

Solution : Integration of Cognitive Computing Services with Object Storage for auto-tagging of unstructured data in the form of objects.

How it works



- IBM has open sourced a sample code that will allow you to auto tag objects in form of images (hosted over IBM Spectrum Scale Object)
- Solution can be extended to other object types depending on business need
- The code and instruction are available on GitHub under Apache 3.0 license
  https://github.com/SpectrumScale/watson-spectrum-scale-object-integration

Thank You