SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2016

Object Storage Analytics: Leveraging Cognitive Computing For Deriving Insights And Relationship

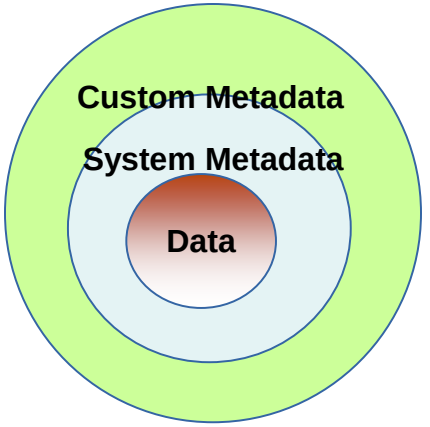
Presenter: Pushkaraj Thorat
IBM India Private Limited

Contributors:
Sandeep Patil, IBM India Private Limited
Nisarga Lolage, IBM India Private Limited

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

Introduction to Object Store



Data + Metadata = Object

- Data is stored as individual objects with unique identifier
- Typically, Objects consist of an object identifier (OID), data and metadata
- Object data is unstructured – images, text, audio, video
- Metadata consists on system metadata and user defined custom metadata that can be extensive

System Metadata

- Filename: taj1234.jpg
- Created: 01 Aug 2016
- Last Modified: 03 Aug 2016

Custom Metadata

- Subject: Taj Mahal
- Place taken: India
- Category: Travel
- Allow Sharing: yes

Object Metadata – Usage

One can assign Object metadata such as

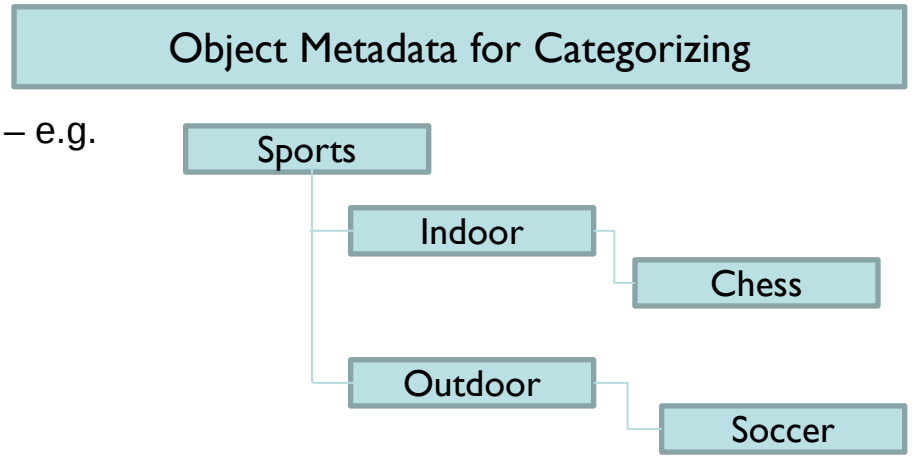
- Tags indicating the contents of the object and type of application the object is associated with.
- The level of data protection / ACLs, Replication / Deletion controlling of object, Movement of object to a different tier of storage/geography;
- The possibilities are limitless.

Indexing/Searching

- Metadata tags are used to categorize data
- Example: Object repository for Sports –

e.g. sport, indoor sport, chess

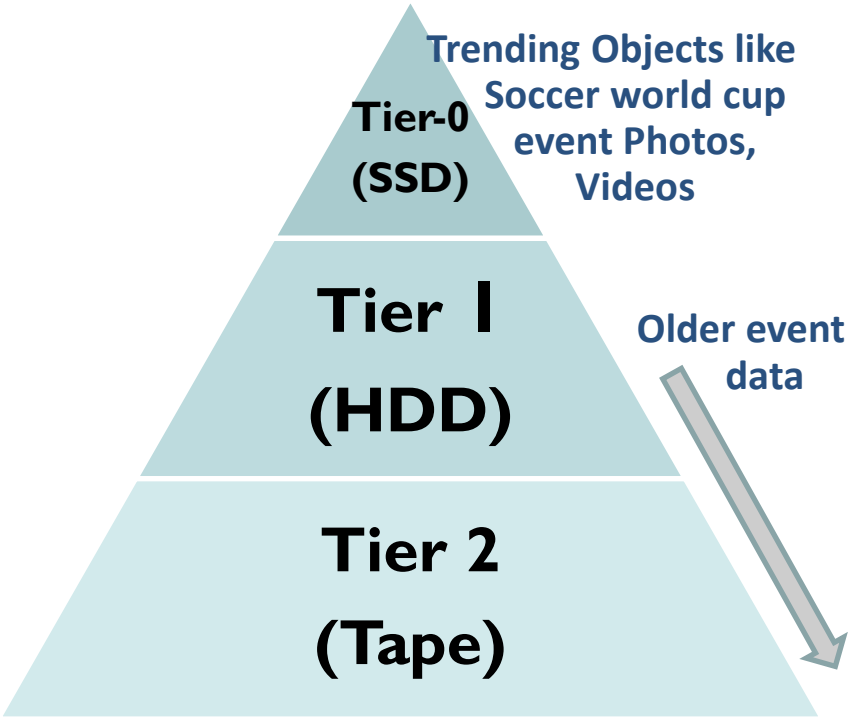
- Objects are then searched based on category – e.g.
- Search All articles related to outdoor sports
- Search Images of all indoor sports
- Search videos related to Chess



Object Metadata – Usage (continued…)

Smart Tiering

- Objects are placed in different tiers of storage pools based on metadata tags
- Allow objects of different categories to be placed in different tiers based on needs
  - Example: Its time for the soccer world cup where objects related to soccer will be potentially highly accessed.
  - Place all the soccer related objects on faster tier.
- Allow independent tiering of objects within the same category
  - Sports Analysts wants to run analytics over only Indoor game. This means one needs to run an Hadoop job on this data.
  - Within “Sport” Category tier only objects tagged as “Indoor” to faster storage pool for analytics.



Analytics and Object Insights

Now we know Object Metadata is very valuable and it's usages are limitless….

But then, Where is the Problem?

Meaningful Tagging of Objects with Metadata ?

For Leveraging the power of User Defined Metadata associated with object, the object has to be appropriately tagged, else it is of less use.

Following are Inhibitors of meaningful tagging of object metadata

- Typical metadata generation processes are
  - **Device-based** (e.g.: Camera tagging basic info to pics)
    - Most of the times these attributes are primitive and low level in nature and provides raw data about that object.
    - Since the tags are Limited or specific they are of less or constrained value for analytics
  - **Manual** – given by user or applications
    - User / applications defined metadata is sometimes looked as unnecessary overhead by the end user at time of object generation and do not tag the data.
    - User might add unnecessary or misleading metadata attributes, which are not adding value for further processing.
- There could be many dimensions of a object, and not all may be added when the object is generated. This too constraints its value. e.g.
  - Image may have many faces, but user might tag only few.
  - A song might be fusion of two genres but in metadata only one is captured.

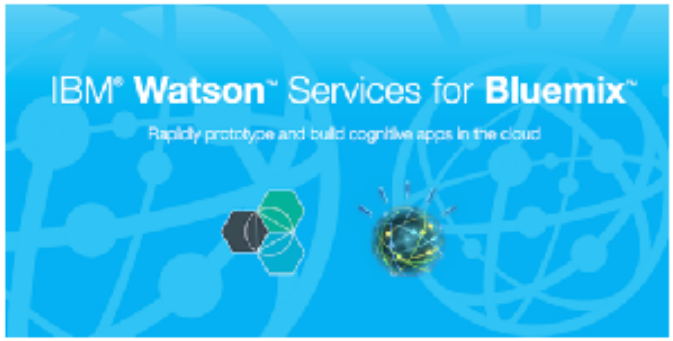
Need for Objects to be accurately auto tagged by Object Stores !

First of a Kind

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

IBM Spectrum Scale
Object Storage

+



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

What is IBM Cognitive Computing 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.

Watson Services

- Alchemy Language – Text Analysis to give Sentiments of the Document
- Language Translation – Translate and publish content in multiple languages
- Tone Analyzer – Discover, understand, and revise the language tones in text
- Visual Recognition – Understand the contents of images.
- 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 …

Example of IBM Watson Cognitive Service

- **Visual Recognition Service**
  - Allows users to understand the contents of an image or video frame.
  - Provides answer to "What is in this image?"
  - Result is in scores for relevant classifiers representing things such as objects, events and settings.
  - e.g. dog (relevance 0.7), mountain (relevance 0.5)

Input Image

Watson Results

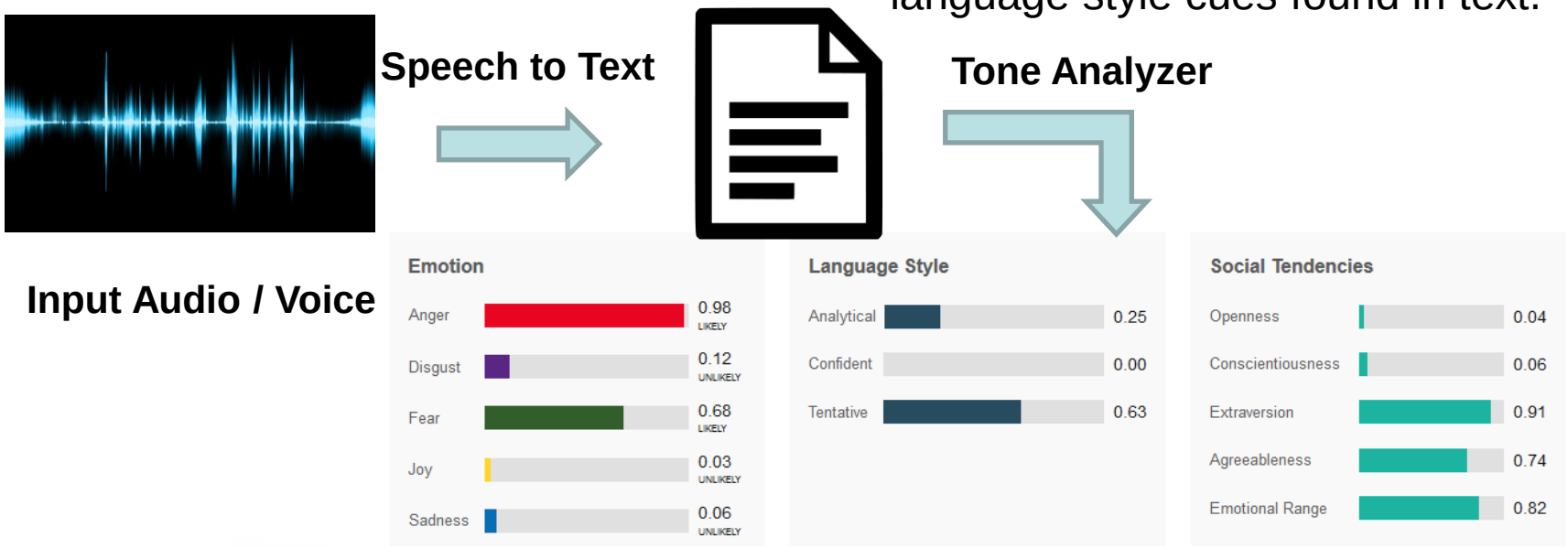
Introduction to Object Store

- Speech to Text

  The Speech to Text service converts the human voice/audio into the written word.

- Tone Analyzer

  This service uses linguistic analysis to detect and interpret emotions, social tendencies, and language style cues found in text.



How Does Cognitive Insights solve the Problem of lack of meaningful Object metadata tags.

- Cognitive Services post deep analysis gives insights of unstructured data (images, audio, video, text, etc.)
- These insights directly relate to the content of the unstructured data and can be used for:
  - Further Analytics of the data
  - Categorization of the data and subsequent use of the categorization in different use cases.
  - Index & Search of data, etc.
- When unstructured data is stored in form of objects, these cognitive insights can be defined and stored as user defined metadata (tags) for the objects:
  - Helps address the problem of meaningful tagging of objects.
  - Opens a realm of newer possibility and opportunity for analytics.

Deriving Object Tags Using Cognitive Services

Cognitive services are asynchronously run by the Object Store to auto-tag the objects.

- Example:
- Visual recognition service is used for images to tag and categorize the images
- Alchemy API's entities and concept tagging for text objects like blogs and text feeds
- Speech-to-text conversion in conjunction with Alchemy API for audio/video objects
- Tone Analyzer can be used over audio files to tag likewise them

Cognitive services categorize the objects into specific group categories.

- Example:
- Animal, canine, tiger
- Sport, outdoor sport, football

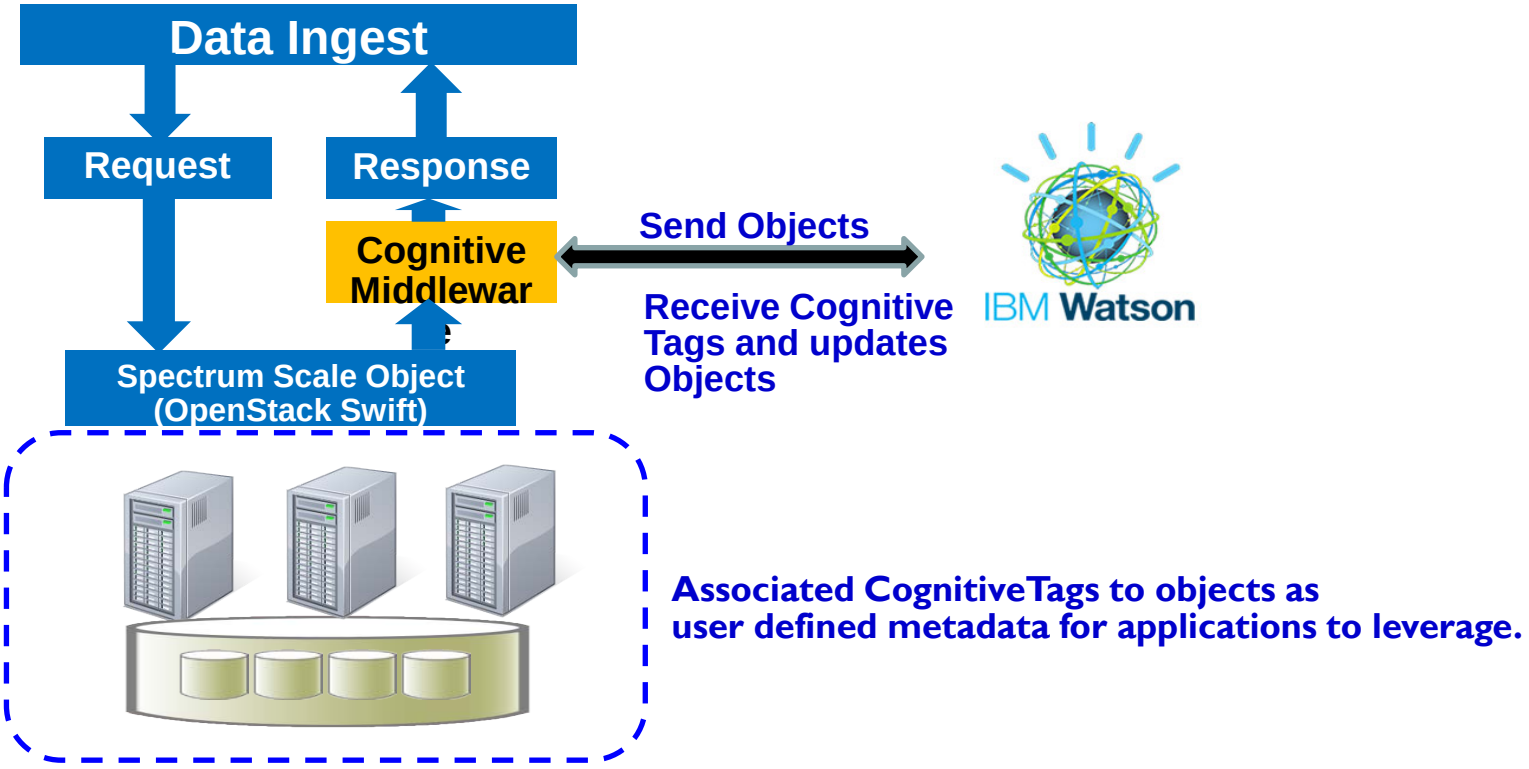
Based on the categories, object relations can be derived, such as

- All articles on indoor sports
- All images of animals
- All nature songs

What Makes it Possible: OpenStack Swift Middleware Framework

- OpenStack Swift is based on WSGI specifications and Middleware is a WSGI feature which extends functionality of any WSGI application.
- Middlewares are heavily used in swift, for purposes such as logging, tempurl, tempauth, quotas etc.
- If there is a middleware in an application pipeline, every request and response is passed through the middleware when a request is served.
- Can I write custom middleware for Spectrum Scale Object ?
  - Yes (needs to be reviewed by development before deployment)

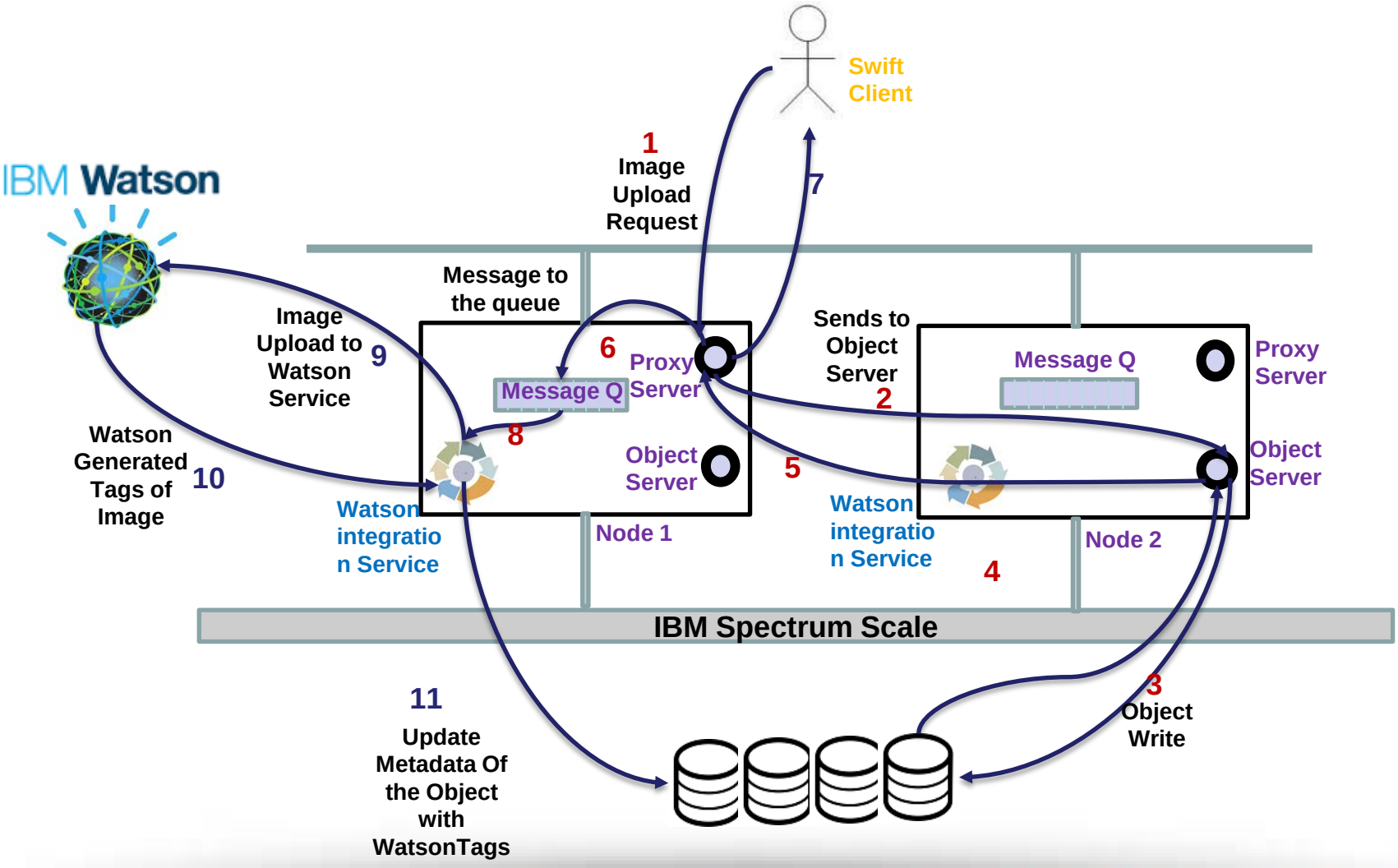
High Level Flow



How OpenStack Swift is integrated with Spectrum Scale

- At very high level Spectrum Scale cluster contains nodes which share common filesystem namespace which is cross mounted on all of its nodes.
- OpenStack Swift cluster consist of multiple type of processes, of which object, container and account servers are backend and Proxy server acts as interface for the cluster.
- There are few designated nodes in Spectrum Scale cluster which are known as ‘Protocol Nodes’ which hosts protocol stack. OpenStack Swift is installed on these protocol nodes.
- All the OpenStack Swift processes are installed on every protocol node of the cluster.
- Depending on the request type i.e. Object or Container or Account, Proxy server chooses the responsible backend server which will serve depending on distributed circular hash, called as Ring.
- So a proxy server can contact any backend server running on other protocol node to fulfill the request.

How Does it Work ?



Design questions and approaches

- At what point should an object sent Watson Cognitive service for analysis.
- Association of IBM Bluemix account used to analyze the object should be associated with cluster or account.

At what point should an object sent Watson Cognitive service for analysis.

- **Approach 1: Uploading the object when user uploads it (middleware based solution)**
- When the user is uploading an object, on successful upload, proxy-server middleware will intercept the request and upload it to Watson Cognitive service
- Metadata will be updated with the results given by the Watson Cognitive Service.

Issues

- Threshold will be reduced.
- Failure to analyze an object will result into (partial) failure of object upload request, also failed object will not be tracked..

- **Approach 2: Batch mode processing**
- Get list of files from file system scans (e.g. os.walk or IBM Spectrum Scale ILM policy scan)
- Identify objects which are not processed by Watson Cognitive Service and process them.

Issues

- File system scans are resource intensive.
- Scans need to be run periodically, hence metadata of newly added objects is not updated immediately.

At what point should an object sent Watson Cognitive service for analysis.

Approach 3: Record the path of object which are newly uploaded, and process them offline (middleware based solution)

- When the user uploads an object, append the object path in a locally running queue, using proxy middleware
- There will be separate service which sends the object to Watson Cognitive Service for analysis and updates metadata.

Issues

- Need mechanism to store the object path.

Currently chosen approach

Association of IBM Bluemix account

Background:

- When one provides swift object store as a service. The ‘account’ entity of Swift is associated to one billing account. This can have multiple users, and users can have roles in that account.
- When we integrate IBM Bluemix thru which Watson Cognitive services are provided, we need to attach this to either particular account or keep it common across whole cluster.
- This is not exactly a technical decision, but a business/deployment one. Answer to this question depends on who are users of the clusters.

Association of IBM Bluemix account

Approach 1: Cluster wide bluemix account

- There would be only one Bluemix account used cluster wide, it will be common across all the accounts.
- Billing will be common for whole cluster.

Issues

- No provision to separate out account/container level billing.

Currently chosen approach

Approach 2: Bluemix account per account

- User would add the bluemix credentials for the account.
- For objects belonging to that account, system will choose these credentials for processing.

Issues

- For every account/tenant in keystone, needs to create an bluemix account and configure it, increasing maintenance overhead.

Final approach

Middleware

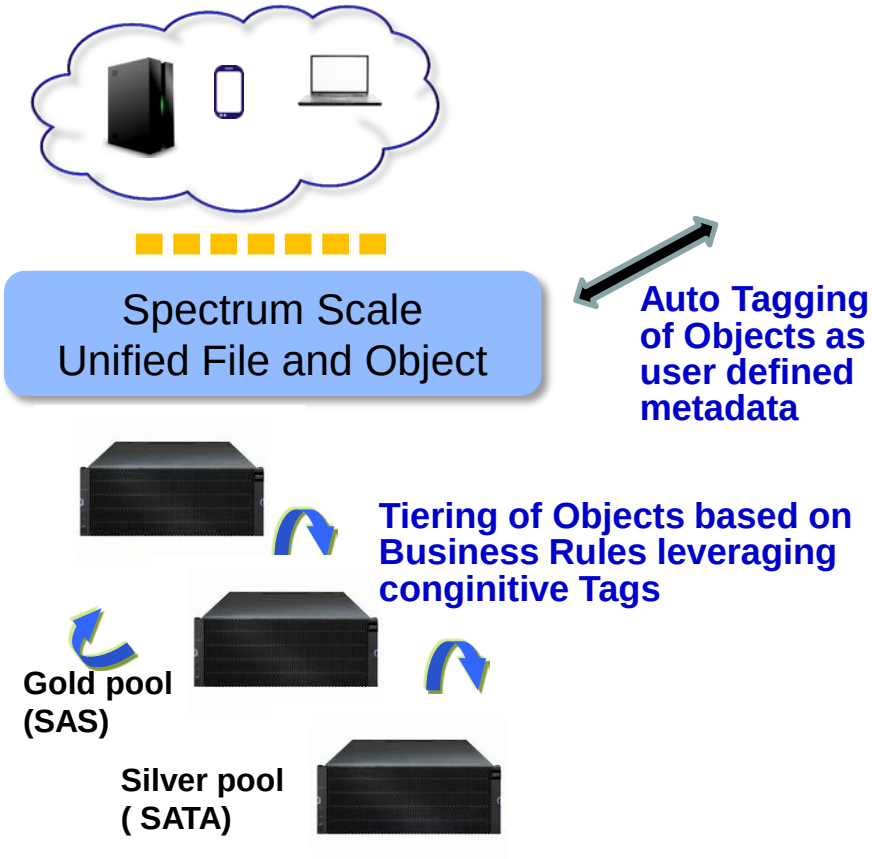
- User will upload the object header – X-Visual_Insights_Enabled to enable it for cognitive analysis.
- Watson Cognitive middleware present in proxy server pipeline will intercept the request in return path, if successful.
- Middleware appends the object path and object timestamp in local queue.

Watson Service

- Is running on every protocol node, and monitors the queue configuration provided.
- When a STOMP message is added in queue by middleware. Watson service checks the timestamp of the message with the file's. If the queue timestamp is older than the file, it ignores the request.
- If the message timestamp matches with the file, it processes the object with Watson Cognitive service and updates the metadata.

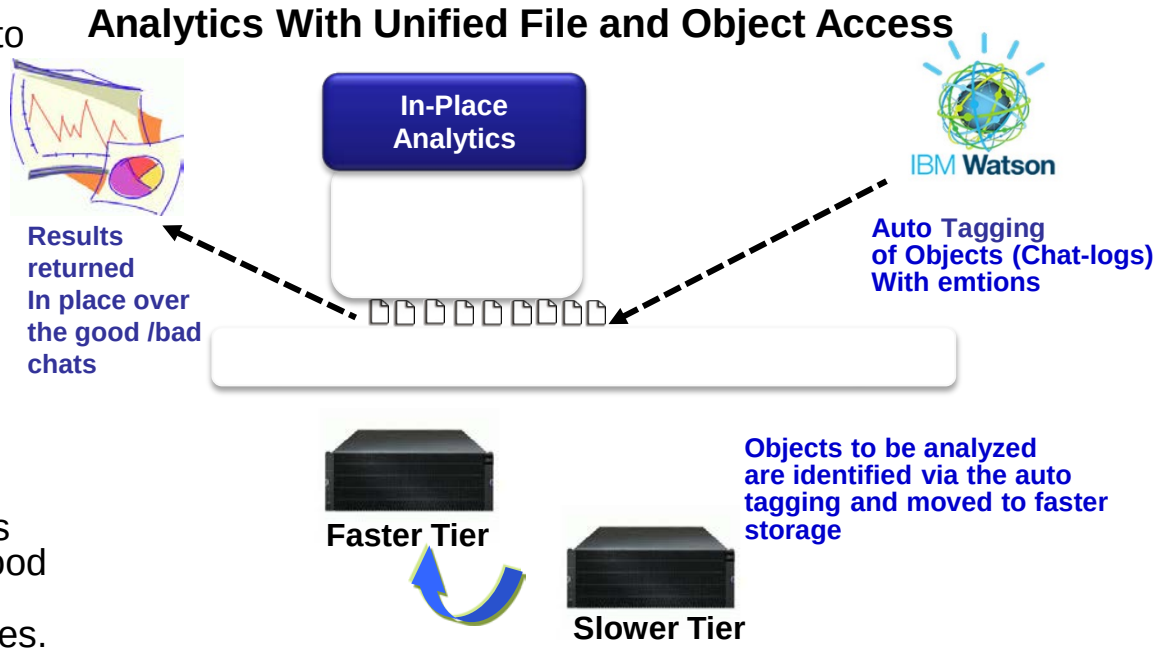
Use Case1: Smart Object Tiering based on Cognitive Tags

- Objects are tagged by cognitive services based on its content.
- Tags can be accessed as xattr by Spectrum Scale placement and migration polices.
- Admins can write placement /movement rules for object based on cognitive tags associated with objects as user defined metadata.
- Example: A sports media portal host sports images and videos for its end users. Assuming Soccer world cup is coming in a month , its end users will access more soccer related content which the portal would like to serve with better response time. With cognitive object tagging on spectrum scale, one can move all images tagged as soccer by the cognitive service to fast tier for better response time for end users.



Use case 2: Chat Center Interaction Analysis

- On-going chat Center Analysis is key to any Business where the business is required to know:
  - % of good chats
  - % of bad chats
  - Analysis over the chat (using Hadoop)
    - Demographic relation to good/bad chat
    - Specific Product relation to good/bad chat
    - etc.
- Chat interaction files stored as Objects are tagged by cognitive services as good & bad interactions or tags based on emotions identified by cognitive services.



- These tags helps categorize data based on type and genre.
- One can then run in-place analytics leveraging spectrum scale Hadoop connectors on the same data to derive more insights.
- Example: Run analytic over all objects marked as “anger emotion chat” to derive the time-line and product being discussed and present a word chart showing which product is generating anger emotion and what were the timelines when it happened.

How to Use this Concept: Its Easy and Simple - Do It Yourself

- We have provided a sample and open sourced the middleware and service code that will allow you to auto tag objects in form of images (hosted over IBM Spectrum Scale Object).
- Based on your business needs and types of objects you can re-use and develop the required middleware
- Spectrum Scale Object which is based on OpenStack Swift supports customer middleware like these to be used, post review with the development.
- The code and instruction are available on GitHub under Apache 3.0 license.
- https://github.com/SpectrumScale/watson-spectrum-scale-object-integration

How to Deploy and Use the Sample Middleware with IBM Spectrum Scale

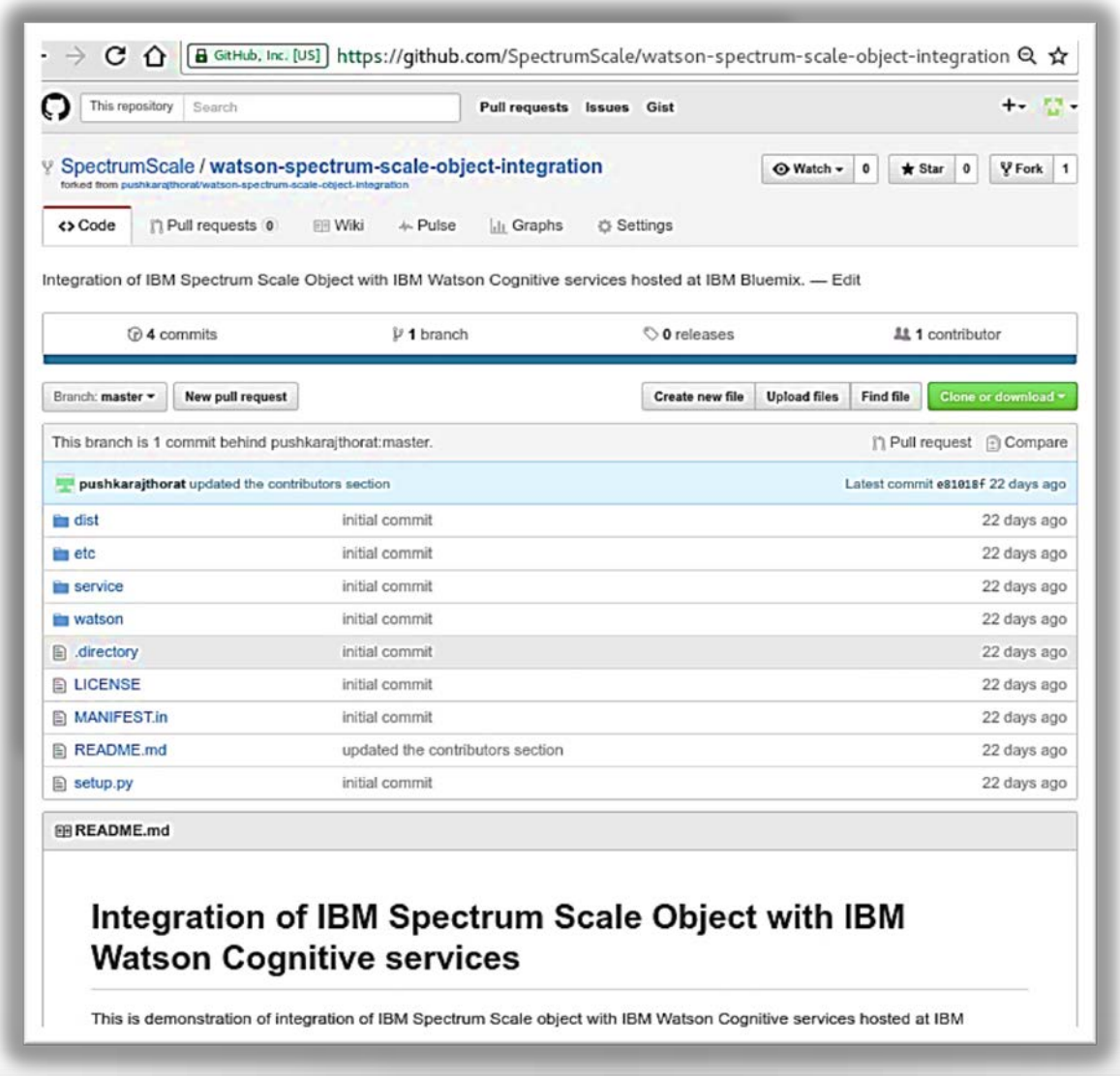
Prerequisite

- Get IBM Bluemix Visual Recognition account
- Install following packages on protocol nodes:
- Stompest https://pypi.python.org/pypi/stompest/
- Apache Active MQ
- Watson Developer Cloud SDK https://pypi.python.org/pypi/watson-developer-cloud
- Ensure connectivity to server - gateway-a.watsonplatform.net from protocol nodes

Deployment

- Install Watson middleware dist/watsonintegration-0.1-1.noarch.rpm on all protocol nodes (from GitHub)
- Update proxy-server.conf to include the middleware, and restart proxy servers.
- Start watsonintegration service on all protocol nodes
- Create a SwiftOnFile policy and create a container with it.
- Upload an image - $ swift upload -H "X-Visual_Insights_Enable:true" <container_name> <object_name>
- Check Watson metadata tags: $ swift stat <container_name> <object_name>

Project location on GITHUB



Future work

- Identification and correction of wrongly tagged objects.
- Making the middleware generic enough to process all types of objects depending on detection of file type.
- Providing flexibility of combining two or more cognitive service together to make extraction of metadata more useful.

Demo

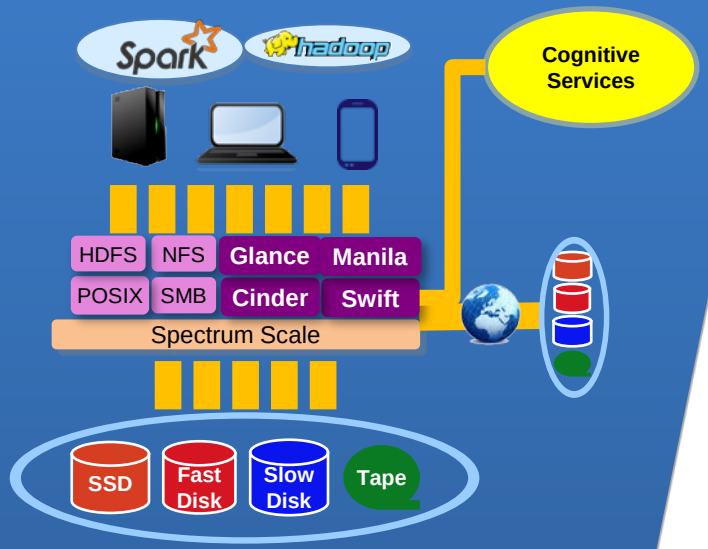
THANK YOU

Acknowledgements

.....

IBM Spectrum Scale

Data management at scale



- Avoid vendor lock-in with true Software Defined Storage and Open Standards
- Seamless performance & capacity scaling
- Automate data management at scale
- Enable global collaboration

OpenStack and Spectrum Scale helps clients manage data at scale

Business: I need virtually unlimited storage	An open & scalable cloud platform
Operations: I need a flexible infrastructure that supports both object and file based storage	A single data plane that supports Cinder, Glance, Swift, Manila as well as NFS, et. al.
Operations: I need to minimize the time it takes to perform common storage management tasks	A fully automated policy based data placement and migration tool
Collaboration: I need to share data between people, departments and sites with low latency.	Sharing with a variety of WAN caching modes

Results

- Converge File and Object based storage under one roof
- Employ enterprise features to protect data, e.g. Snapshots, Backup, and Disaster Recovery
- Support native file, block and object sharing to data.

Notices and Disclaimers

- Copyright © 2016 by International Business Machines Corporation (IBM). No part of this document may be reproduced or transmitted in any form without written permission from IBM.
- **U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**
- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY. IBM products and services are warranted according to the terms and conditions of the agreements under which they are provided.
- IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
- **Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**
- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
- It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law

Notices and Disclaimers Cont.

- Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. IBM EXPRESSLY DISCLAIMS ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

- The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

- IBM, the IBM logo, ibm.com, Aspera®, Bluemix, Blueworks Live, CICS, Clearcase, Cognos®, DOORS®, Emptoris®, Enterprise Document Management System™, FASP®, FileNet®, Global Business Services ®, Global Technology Services ®, IBM ExperienceOne™, IBM SmartCloud®, IBM Social Business®, Information on Demand, ILOG, Maximo®, MQIntegrator®, MQSeries®, Netcool®, OMEGAMON, OpenPower, PureAnalytics™, PureApplication®, pureCluster™, PureCoverage®, PureData®, PureExperience®, PureFlex®, pureQuery®, pureScale®, PureSystems®, QRadar®, Rational®, Rhapsody®, Smarter Commerce®, SoDA, SPSS, Sterling Commerce®, StoredIQ, Tealeaf®, Tivoli®, Trusteer®, Unica®, urban{code}®, Watson, WebSphere®, Worklight®, X-Force® and System z® Z/OS, are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.