VERITAS

illuminate

Shining light on dark data

Machine Learning based Tiering in Access

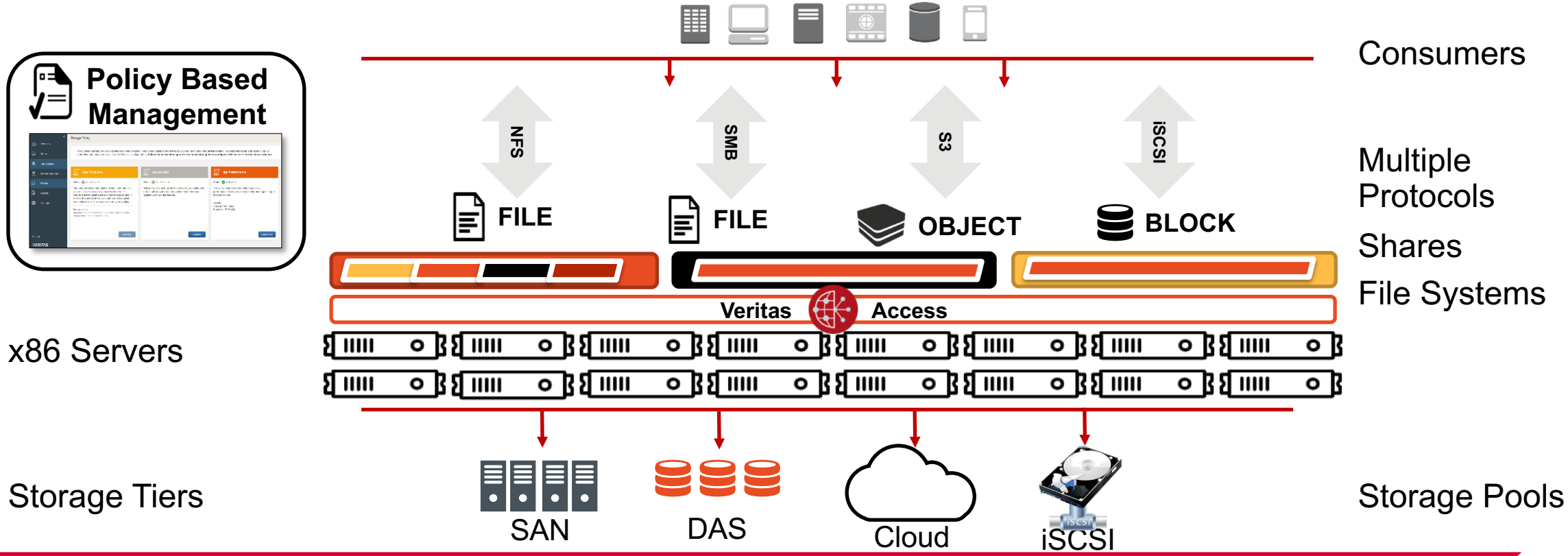
Niranjan Pendharkar, Anindya Banerjee

Agenda

1	What is Veritas Access
2	Why
3	How

Veritas Access

Software-Defined Scale-Out NAS



Agenda

1. What is Veritas Access
2. Why
3. How

Tiering

- Supports multiple tiers
  - Cost
  - Performance
  - Reliability
- What to move
  - Based on policies
- Where to move
  - Again based on policies
- When to move

Tiering policies

- Access temperature (number of accesses)
- IO temperature (bytes read/written)
- Last access time
- Last modification time
- Type of the file
- Size based
- Content of the file
- Tags set on the file

Provision Storage for Backup

1. Storage Options
2. Cloud Options
3. Summary

LTR On-Premises + Cloud

This policy enables Backup images to be saved on multiple storage tiers using on-premises and cloud storage. Users can use this policy to move the backed-up images to cloud that are not modified over a specified period of time. On-premises backup images are highly available. This policy optionally enables protection against device failures.

Capabilities

- Fault Tolerance
- Tiered File System

Storage Pool: **spool** Pool Size: **45 GB**

On premises → Cloud

Cloud Storage Options

Move Images not modified for 10 days

Service Provider AZURE

REST End Point azure.microsoft.com

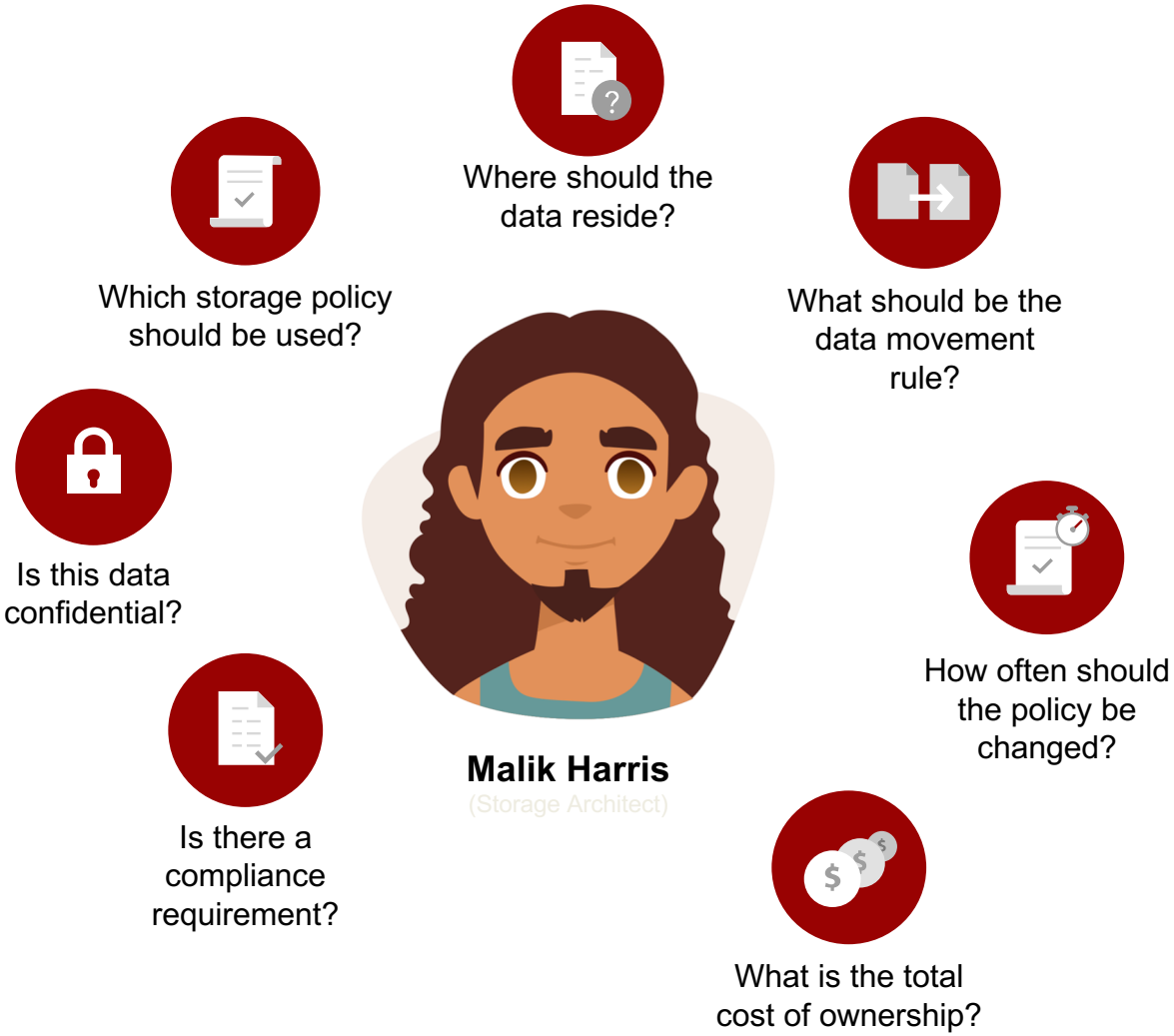
Tier Type azure

Move Images to cloud at 04:00 Hours 00:00 Minutes

Cancel Previous Next

Manual Configuration

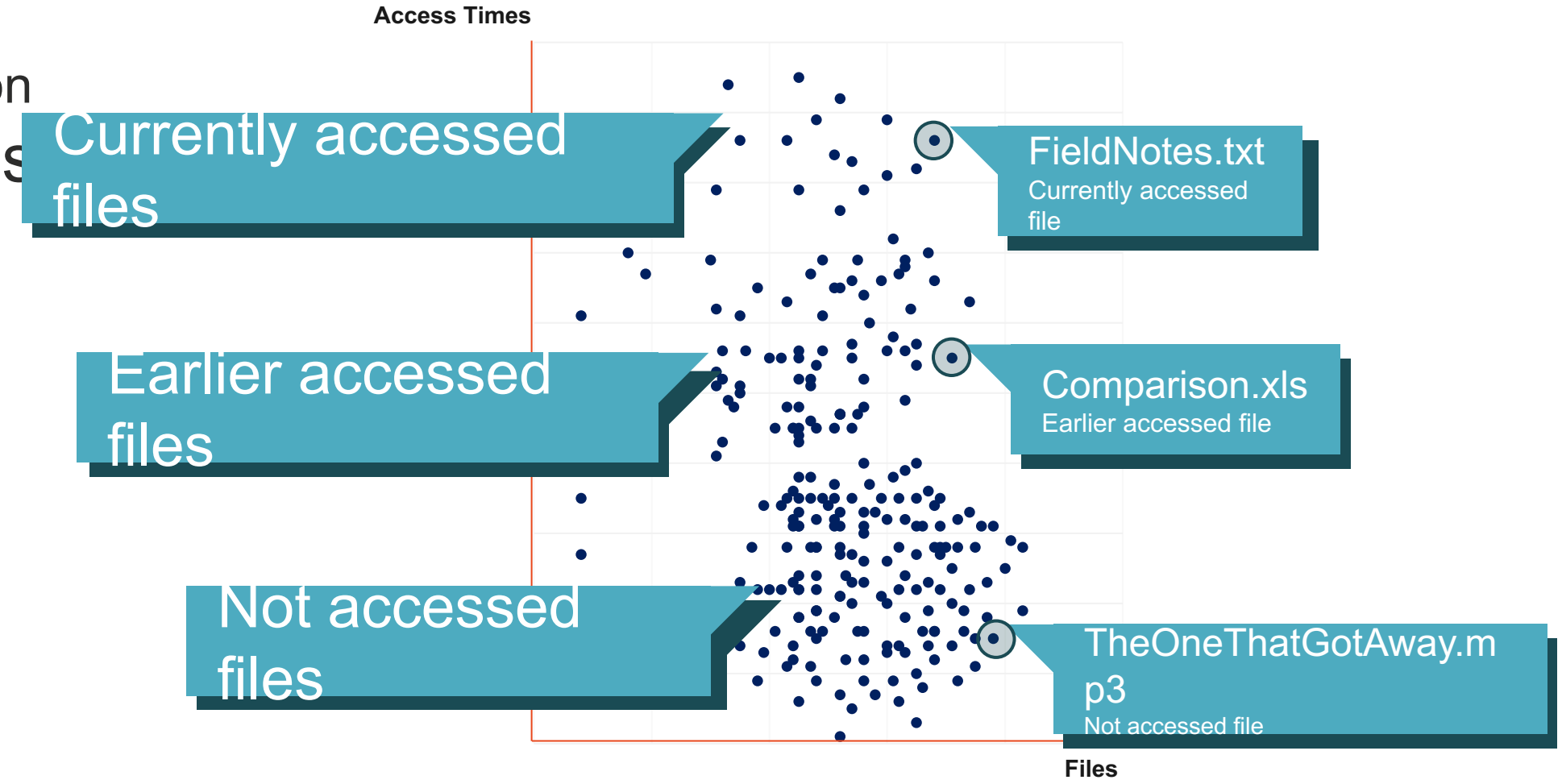
Being an Admin



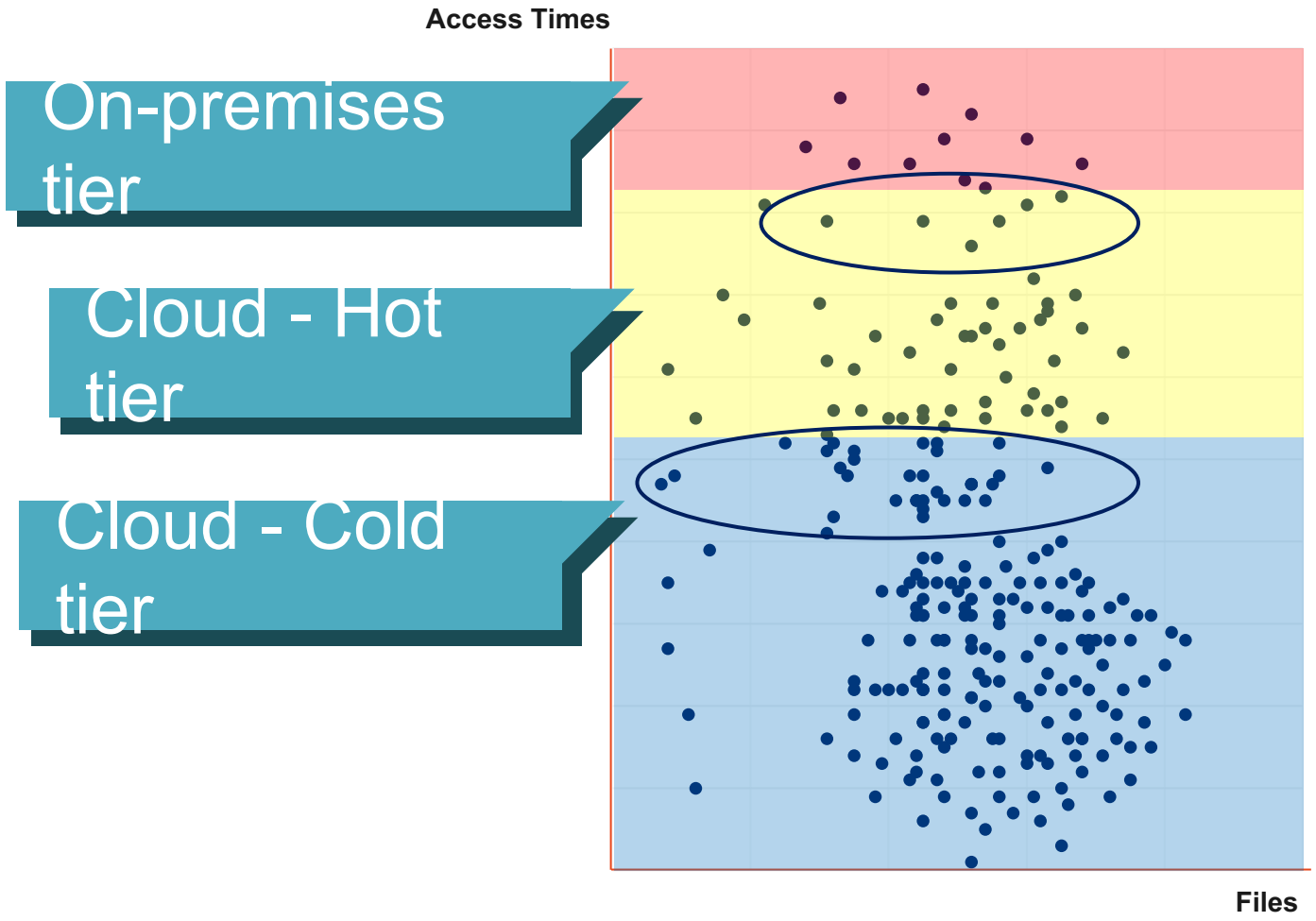
How we decide which file goes to which tier?

Based on

Access



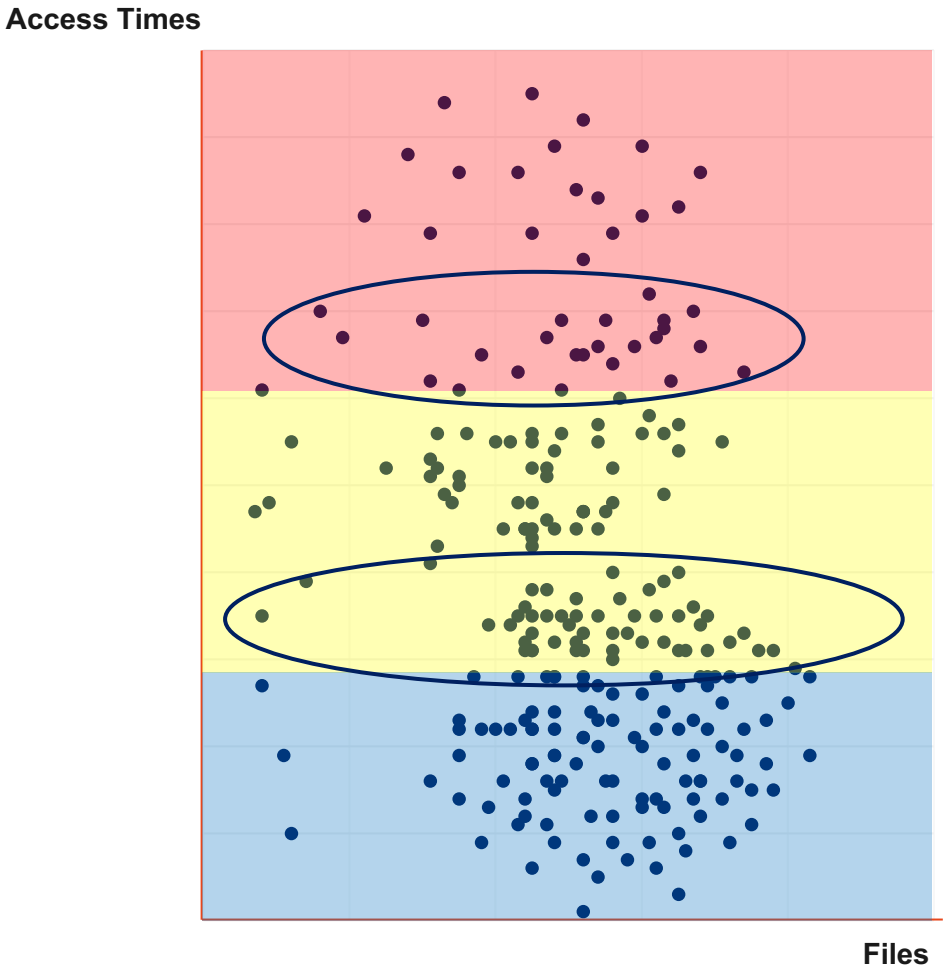
Aggressive Tiering



Tier	Storage required	Cost
On-premises	4 TB	$ 14K
Cloud – Hot	125 TB	$ 115K
Cloud – Cold	380 TB	
Total	509 T	

$ 185K

Conservative Tiering



Tier	Storage required	Cost
On-premises	12 TB	$ 44K
Cloud – Hot	237 TB	$ 218K
Cloud – Cold	260 TB	
Total	509	$ 300K

How wonderful it would be to automatically set this right!

Veritas™ Access 7.3

Quick Actions

Overview
Shares
File Systems
NAS Infrastructure
Policies
Reports
Settings

Contact

VERITAS

Settings / Intelligent Storage Tiering

Intelligent Storage Tiering

Intelligent storage tiering u gorithms on log data to model dynamic storage tiers.
It uses multiple data parame s temperature, access time and classification of data to do the modeling.

Agenda

1 What is Veritas Access

2 Why

3 How

Data sources

- List of all the files
  - Along with their attributes
- File Change Log (FCL)
  - Captures file accesses and modifications
- IMI
  - Captures file accesses and modifications
  - Lightweight compared to FCL
  - More granular
- Various statistics
  - Vxfsstat, vmstat, iostat, sar output …

ML in the works

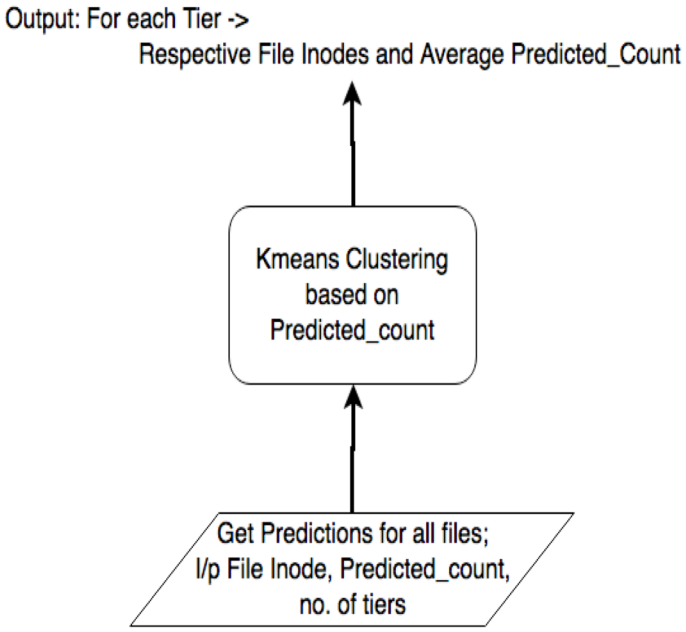
- Data Pre-processing
  - Involves cleaning of data from the stats Access generates
  - Data is prepared for training by choosing only required fields (features)
  - if any data is missing, it is added to maintain consistency
- Model Creation (or Learning)
  - Data prepared in the previous step is sent to ML Algorithms
  - ML Algorithms learn the past File Access patterns and generate a model
- Model Serving
  - Model created in the previous step is used to make future predictions of File Access

ML Algorithms

- Pattern Prediction Algorithm (Unsupervised Approach)
  - Initially ARIMA was tried
  - 5 layered Deep Neural Network
    - combination of Recurrent Neural Network and Dense Neural Network
    - Input – File usage statistics generated by Access/Infoscale
    - Output – Prediction of how likely the file is going to be accessed in the future
  - Also exploring Prophet (from Facebook)
- RNN (Recurrent Neural Network)
  - Makes use of sequential information - good fit for Time Series data
  - Very good at capturing long term dependencies
  - Reinforcement Learning - learns from its mistakes and updates the model accordingly
- DNN (Dense Neural Network)
  - Feedforward network with many hidden layers

ML Algorithms - contd

- k-means Clustering (Unsupervised Approach)
  - Prediction algorithm predicts frequency of a file's access in the future
  - Clustering is done on these values to place files in different tiers

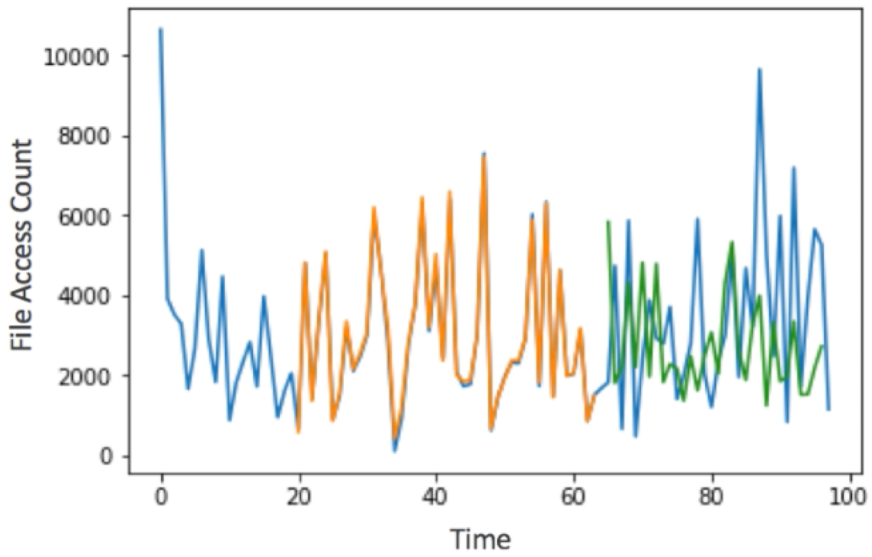


Really?

	Id	Device_ID	OPCODE	User_Name	Domain_Name	SID	Offset	Length	Path	RenamePath	Type	Image_Name	IPAddr	TimeStamp
0	1	1	4	0	0	NaN	1843200.0	1024.0	/mnt1/bstress3/70	NaN	2	NaN	NaN	1493190503
1	1456	1	1	0	0	NaN	NaN	NaN	/mnt1/bstress1 /fsr720-01vm3 /0/101 /012345678901...	NaN	2	NaN	NaN	1493190504
2	1458	1	200000	0	0	NaN	NaN	NaN	/mnt1/bstress1 /fsr720-01vm3 /0/101 /012345678901...	NaN	2	NaN	NaN	1493190504
3	1459	1	18	0	0	NaN	NaN	NaN	/mnt1/bstress1 /fsr720-01vm3 /0/101 /012345678901...	NaN	2	NaN	NaN	1493190504
4	1468	1	18	0	0	NaN	NaN	NaN	/mnt1/bstress1 /fsr720-01vm3 /0/101 /012345678901...	NaN	2	NaN	NaN	1493190504

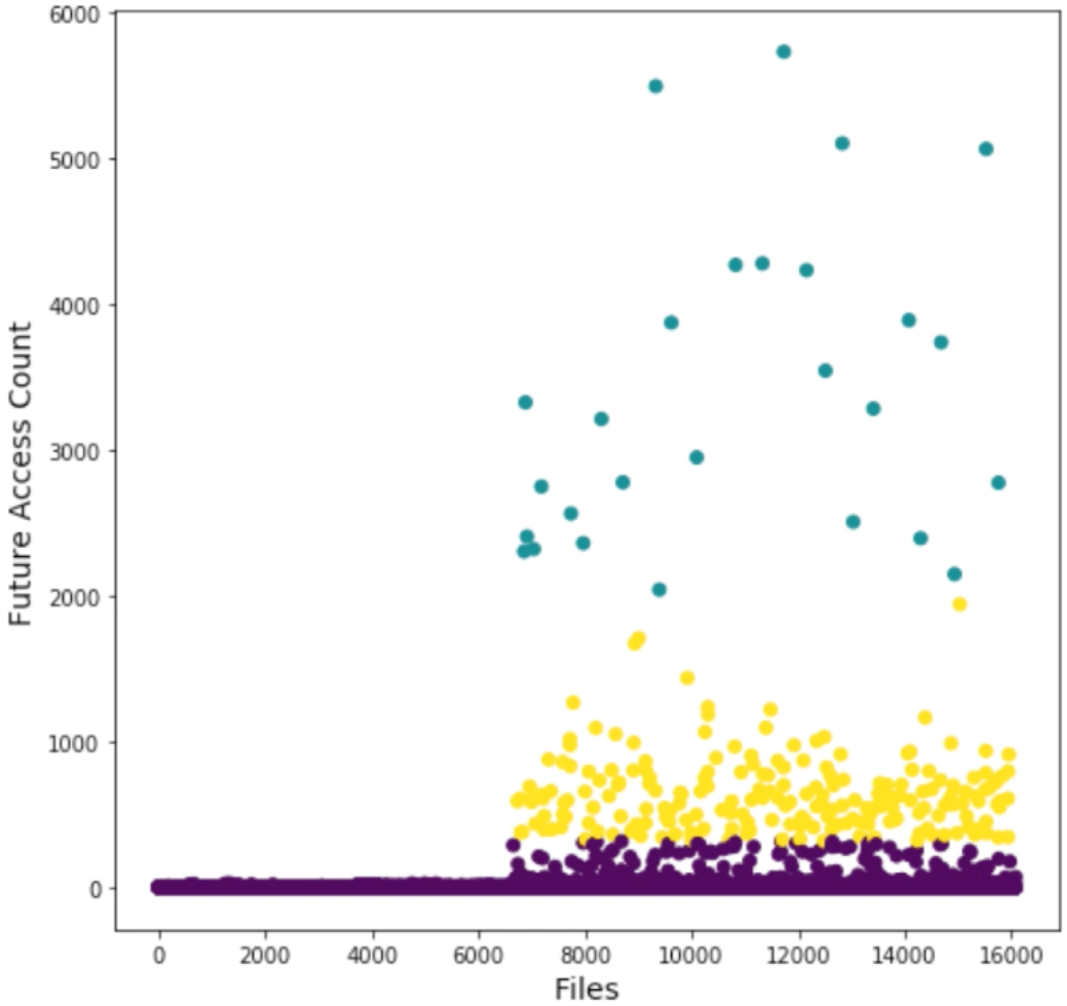
In action

	Inode	Count	Count_0	Count_1	Count_2	Count_3
0	1	1	1	0	0	0
1	2	5	5	0	0	0
2	5	1	1	0	0	0
3	6	5	5	0	0	0
4	9	1	1	0	0	0



	File Name	Future Access Count Prediciton
0	/mnt1/bstress1/fsr720-01vm3/0/101/012345678901...	5
1	/mnt1/bstress1/fsr720-01vm3/0/1013/01234567890...	3
2	/mnt1/bstress1/fsr720-01vm3/0/1017	1
3	/mnt1/bstress1/fsr720-01vm3/0/102	1
4	/mnt1/bstress1/fsr720-01vm3/0/102/012345678901...	5

And finally



Considerations

- Runs periodically
- During each run,
  - Consider model developed in previous run
  - Consider IMI information during the interval
  - Update model accordingly – feedback loop
- Computationally intensive
- Have ML packages as a service in platform
  - Input: IMI information
  - Output: Tunables based on model

Veritas™ Access 7.3

Quick Actions

Intelligent Storage Tiering

1 Level of confidence

2 Tiering Options

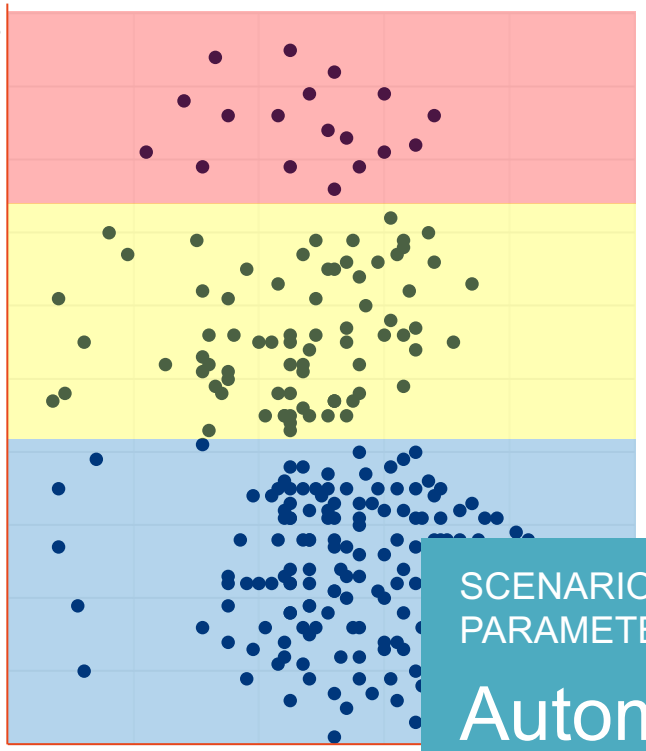
3 Cost of Ownership

Level of confidence

92%

Tiering based on the level of confidence

Access Times



Tier	Storage required	Cost
On-premises	5 TB	$ 18K
Cloud – Hot	63 TB	$ 58K
Cloud – Cold	441 TB	
Total	509 TB	$ 141K

SCENARIO 1: CLUSTERING BASED ON A SINGLE PARAMETER

Automatic, accurate and dynamic tiering

Veritas™ Access 7.3

Quick Actions

Intelligent Storage Tiering

1 Level of confidence

2 Tiering Options

3 Cost of Ownership

Level of confidence

92%

Tiering based on the level of confidence

Confidential or currently accessed data

Non confidential and earlier accessed data

Non confidential and not accessed data

Tier	Storage required	Cost
On-premises	9 TB	$ 33K
Cloud – Hot	62 TB	$ 57K
Cloud – Cold	438 TB	
Total	509 TB	$ 154K

SCENARIO 2: CLUSTERING BASED ON MULTIPLE PARAMETERS

Automatic, accurate and dynamic tiering

Cancel Previous Next

Veritas™ Access 7.3

Quick Actions

Intelligent Storage Tiering

1. ✓ Level of confidence
2. Tiering Options
3. Cost of Ownership

Select cloud subscription for different tiers.

Tier	Storage required	Cloud subscription (for 3 years)		
On-premises	9 TB	On premises storage $ 33K		
Cloud – Hot	62 TB	Cloud Provider 1 $ 57K	Cloud Provider 2 $ 60K	Cloud Provider 3 $ 60K
Cloud – Cold	438 TB	Cloud Provider 1 $ 65K	Cloud Provider 2 $ 72K	Cloud Provider 3 $ 60K

Note: Suggestions for cloud subscription is based on storage and retrieval cost.

Cancel | Previous | Next

Summary

- Storage solutions supports storage tiering
- Machine learning can automate
  - User need not set thresholds
- Predict feature usage using neural networks
- Group items using clustering techniques
- Automatically move items to ideal storage tiers

THANK YOU