AI-based Data Popularity, Placement Optimization for a Tiered Storage architecture at BNL/SDCC Facility

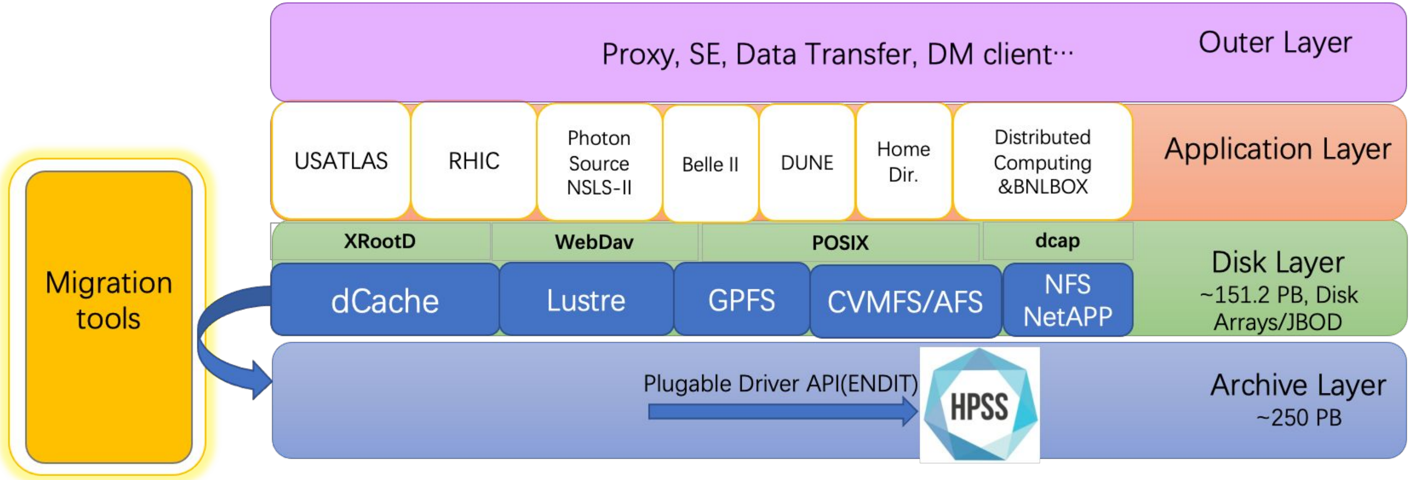
Qiulan Huang, James Leonardi, Carlos Deleon, Vincent Garonne, Shinjae Yoo

Brookhaven National Laboratory

@BrookhavenLab

ACAT 2024, SBU, NewYork - Mar 11 - Mar 15, 2024

Storage Overview at BNL/SDCC

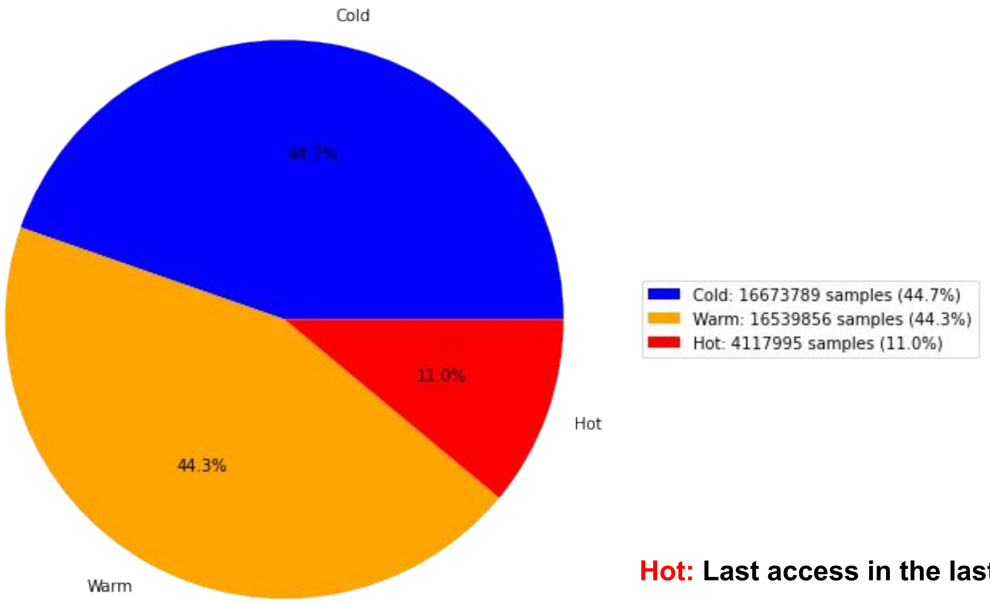


- **Several millions of files created/deleted/transferred per day**
- **Peak traffic at 100GB/s**
- **Data accessed by millions of jobs per day**
- **Hundred of thousands of storage devices**

- Tiered Storage
  - Encompasses various storage technologies to serve different workloads and use cases (HPC posix access, HTC grid access, …)
  - Involve different generations of storage over a period

Data Temperature（Take ATLAS data for example）

Jan 1, 2023-Dec 31, 2023, ~37 million files



Hot: Last access in the last month

Warm: Last access in the last 6 months

Cold: Last access between 6 months and one year

AI/ML For Storage Optimization

Motivation

- In the current tiered storage "class” system at the Data Center
  - Unused data is stored on expensive storage
  - Fast IO storage is not currently used effectively

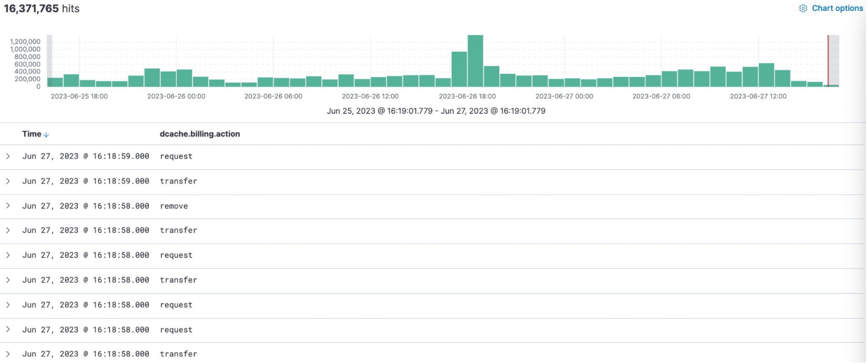
Goals

- Design an efficient monitoring platform to collect the relevant information from various distributed data sources
- Develop an optimal data management system for the data center to maximize usable space while minimizing access latency, within budget, hardware, and compliance constraints
  - Heavy use of storage, metadata and data popularity information
  - Develop a precise AI/ML prediction model to possibly forecast the future usage of the data
  - Orchestration of data for optimal movement and placement

Data collection

- Has collected data of the past 2 years
  - Data volume: ~11TB
  - **~10GB** in average per day, **5~8 million events** per day
  - Data source: billing logs, domain logs, etc from various experiments like usatlas, Belle2, etc

Time: one day	size	records
Raw data	13GB	5,604,498
Preprocessed data	2.7GB	5,604,498



Data preprocessing

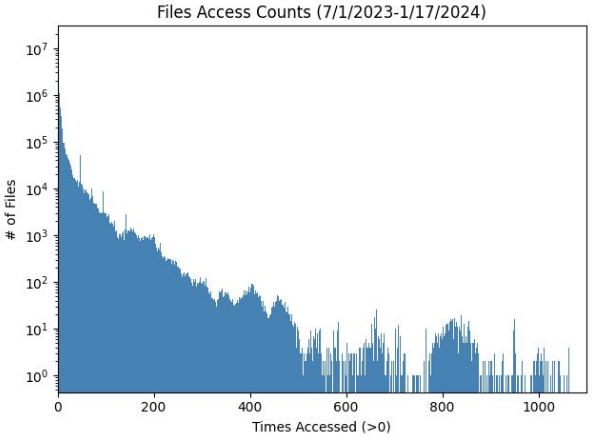
- Define and generate the tabular data or comma-separated values (CSV) file format for data training and facilitates finding patterns between files
  - pnfsid
  - Access Count
  - Access Timestamps
  - Rucio Scope (mc15_13TeV)
  - Task ID
  - Datatype (DAOD, EVNT, HIST, etc.)
  - Avg Time Between Accesses
  - Action(create, transfer, delete,)
  - User ID
  - …

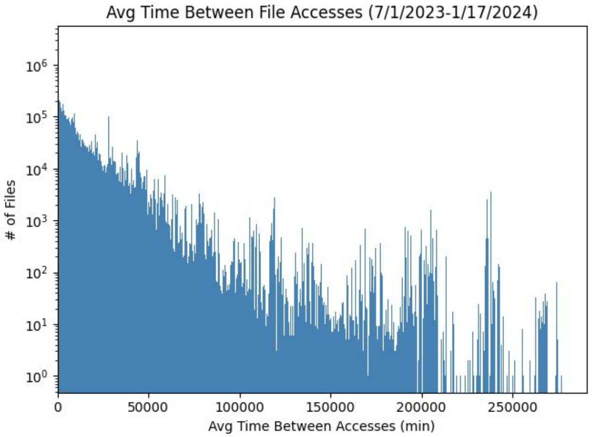
File ID	path	taskid	datatype	scope	First_Access	Last_Access	…
file_1							
file_2							
…							

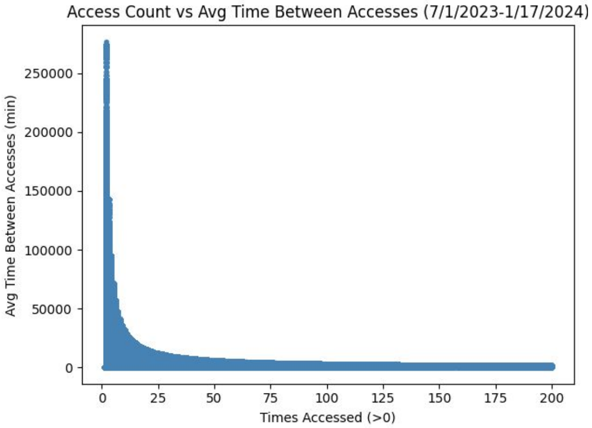

```
pnfsid|path|taskid|datatype|scope|accesscount|clientips|protocols|actions|firstaccess|accesstimes|lastaccess|mintimebetween|avgtimebetween|maxtimebetween|errorcodes
0000A3EECFE022224142A68A0037FE3A446D|/pnfs/usatlas.bnl.gov/BNLT0D1/rucio/mc23_13p6TeV/d6/ad/DAOD_PHYSLITE.35040159._000250.pool.root.1|35040159|DAOD_PHYSLITE|mc23_13p6TeV|1|{'130.199.206.137'}|{'Xro
otd-5.0'}|{'request'}|2023-11-01 00:00:02.540000-0400|{'2023-11-01 00:00:02.540000-0400'}|2023-11-01 00:00:02.540000-0400|0|0|0|{'0'}
00008583BBF8DD8A4B0787679565564E2794|/pnfs/usatlas.bnl.gov/BNLT0D1/rucio/mc23_13p6TeV/1e/d6/DAOD_PHYSLITE.35040159._000342.pool.root.1|35040159|DAOD_PHYSLITE|mc23_13p6TeV|1|{'130.199.206.149'}|{'Xro
otd-5.0'}|{'request'}|2023-11-01 00:00:05.428000-0400|{'2023-11-01 00:00:05.428000-0400'}|2023-11-01 00:00:05.428000-0400|0|0|0|{'0'}
000058B7CD9318E44138857679E39F0E5B17|/pnfs/usatlas.bnl.gov/BNLT0D1/rucio/mc23_13p6TeV/a4/60/DAOD_PHYSLITE.35040159._000330.pool.root.1|35040159|DAOD_PHYSLITE|mc23_13p6TeV|1|{'130.199.156.199'}|{'Xro
otd-5.0'}|{'request'}|2023-11-01 00:00:06.400000-0400|{'2023-11-01 00:00:06.400000-0400'}|2023-11-01 00:00:06.400000-0400|0|0|0|{'0'}
000058BC8CE6F325496B982EF0ABF2B2AF05|/pnfs/usatlas.bnl.gov/BNLT0D1/rucio/mc23_13p6TeV/7d/9c/DAOD_PHYSLITE.35040159._000253.pool.root.1|35040159|DAOD_PHYSLITE|mc23_13p6TeV|1|{'130.199.159.140'}|{'Xro
otd-5.0'}|{'request'}|2023-11-01 00:00:06.777000-0400|{'2023-11-01 00:00:06.777000-0400'}|2023-11-01 00:00:06.777000-0400|0|0|0|{'0'}
0000223C108F5ED14EB59CAA13263B97E30F|/pnfs/usatlas.bnl.gov/BNLT0D1/rucio/mc20_13TeV/01/97/AOD.35261114._000644.pool.root.1|35261114|AOD|mc20_13TeV|2|{'130.199.206.204'}|{'Xrootd-5.0'}|{'request'}|20
23-11-01 00:00:07.714000-0400|{'2023-11-01 00:00:07.714000-0400', '2023-11-01 00:00:07.757000-0400'}|2023-11-01 00:00:07.757000-0400|0.043|0.043|0.043|{'0'}
```


Brookhaven National Laboratory

Data Analysis- Access Distribution







- Majority of files accessed less than 200 times
- As files are accessed more, time between accesses tends to decrease
- Rightmost plot trimmed to show patterns

Brookhaven National Laboratory

Exploring Data Correlation

- Since we predict the data popularity in the future, it will be useful to know which files are accessed with each other
    - If one file is accessed, this can push other files to become 'hot' as well.
- Goal: Generate a Co-Occurrence Matrix
    - Visualize which files are accessed with each other.
- For figure on right
    - Each number represents a different file
    - Put all files along each axis
    - Count how many times 1 followed by 0

a

3	1	1	0	1
0	1	2	2	1
1	0	1	2	1
3	1	3	1	3
0	1	1	3	1

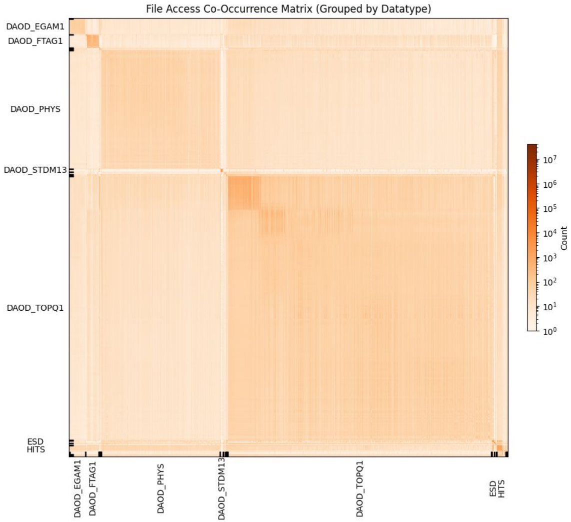
b

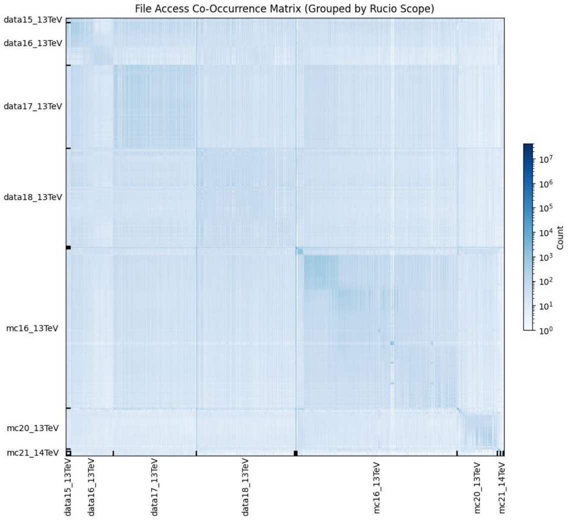
Pixel pair	0	1	2	3
0	0	4	0	0
1	2	2	2	3
2	0	2	1	0
3	0	4	0	0

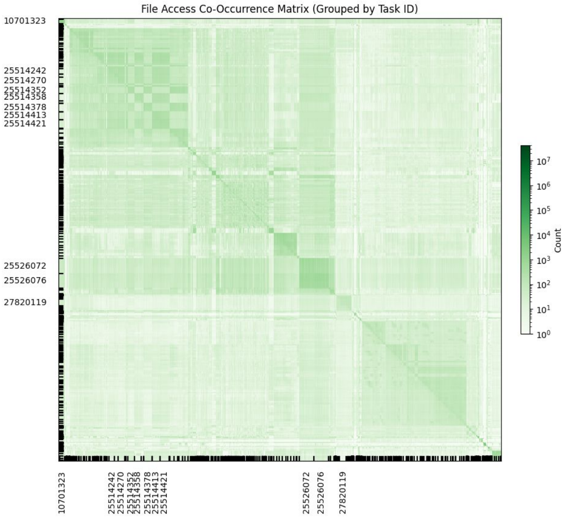
Example of co-occurrence matrix.
Source:https://www.researchgate.net/figure/Gray-level-co-occurrence-matrix-calculation-example-For-interpretation-of-the_fig5_273731213

The Data Co-occurrence Matrix

- ➔ Group by any desired attribute: Task ID, Rucio Scope, Datatype, etc.
- ➔ Patterns appear along diagonal
- ➔ Denote highly correlated groupings





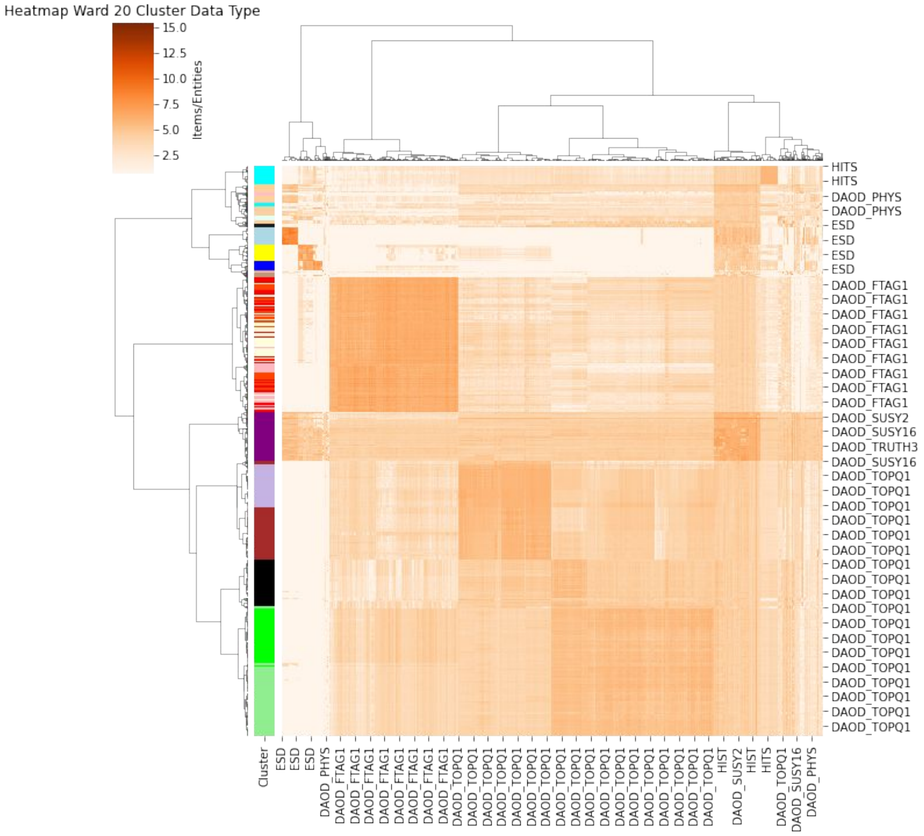


- Matrices are expensive (quadratic time + space complexity)
- Focus on highly-accessed files (150+ access times, 90K files)
  - Likely to be accessed again
  - The matrix size reduce from 23 million×23 million to 90K×90K

Brookhaven National Laboratory

Data Analysis- Clustering

- Perform unsupervised learning
- Explore the patterns that help to differentiate the data
- A clear pattern in data type shown in the matrix correlation as well as the dendrogram hierarchical clustering and the K-means clustering
- All 3 clustering methods show a pattern is connected to the datatype feature



Brookhaven National Laboratory

Data Training

- Data samples: 6 months data (~23 million files)
- **Features**:hold patterns that were shown in previous slide

  ['taskid', 'datatype', 'scope', 'accesscount', 'avgtimebetween']
- **Feature importance**

  taskid features: 0.4534

  avgtimebetween features: 0.1404

  accesscount features: 0.1066

  datatype features: 0.2193

  scope features: 0.0803

  Sum of importances for features: 1.0000

➔ The features we used to train our model all impact the model differently. Some of our features impact the model more than others. The % of each feature tells us how much of an impact it is to the decision tree when determining the classification

Prediction Model and Results

Model Architecture:

- Input of the model: one-hot encoding of the Categorical columns
- Output of the model: hot/warm/cold classification

Model Training:

- Features: [‘taskid’, 'datatype', 'scope', 'accesscount', 'avgtimebetween']
- Labeled data temperature based on the last accessed file which we removed from the training
- Randomly selected 60k samples to use for model training
    - 20k samples for each Hot, Warm, Cold 12k for validation(4k each type) and 48k for training

Results(More details see the the backup slides 19-23):

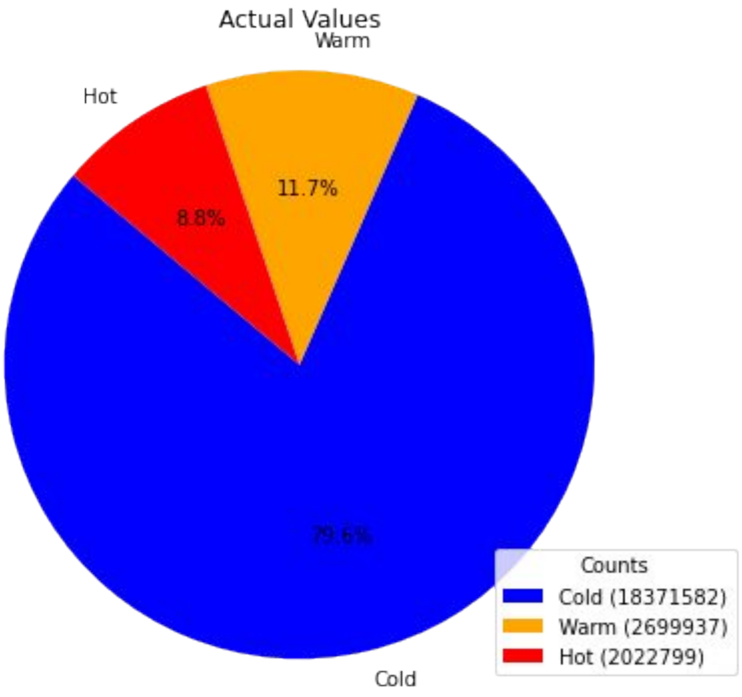
- The model's performance is evaluated on the different sets to assess its predictive accuracy, precision, and recall
- With the larger dataset, the accuracy improves, highlighting the benefits of increased training data
- Precision improves with the more even # of each type(hot/warm/cold)

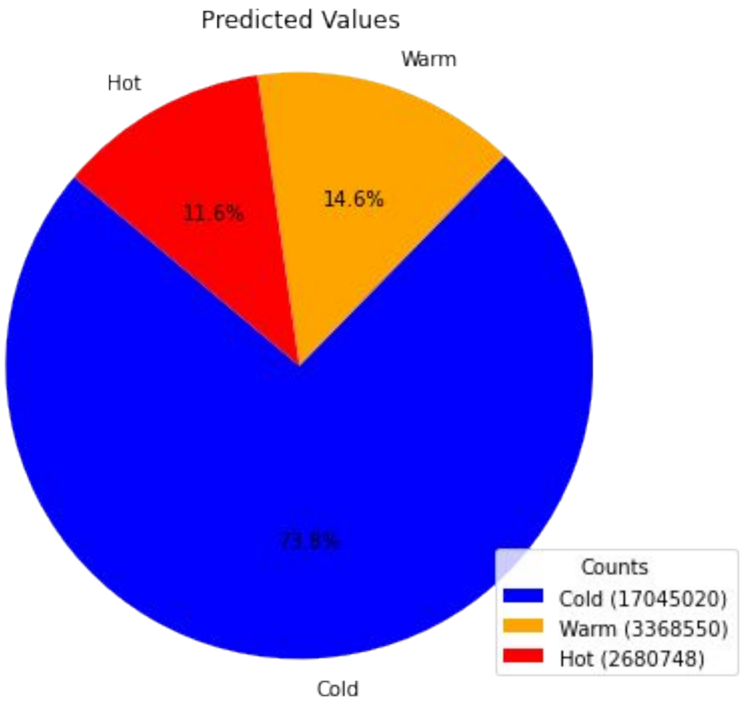
	Set 1 (Initial 60K)	Set 2 (Top 300K)	Set 3 (Total 23M)	Set 4 (Random 1.5M)	Set 5 (Random 1.5M, Even # of each type)
Accuracy	91.68%	90.70%	**91.81%**	90.40%	90.86%
Recall	91.66%	92.00%	91.33%	91.66%	91.00%
Precision	91.66%	82.33%	80.33%	74.33%	91.33%

Brookhaven National Laboratory

Labeled vs Prediction popularity

Hot: 0-1 Week
Warm:1 week - 3 months
Cold: 3+ months

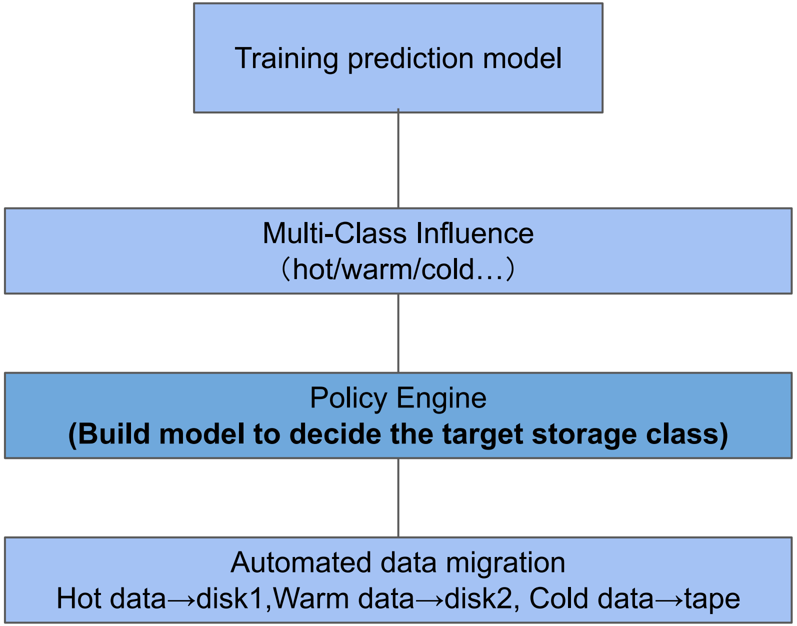




Policy engine

- The objective is to propose and evaluate data migration strategies for optimizing data storage
- The input data output(y)=input(x), y contains {hot, warm, cold}
- Build a model to decide the target storage class for data migration
  - Metrics: user response time, load, CPU, disk space,etc
  - Define different weights for the metrics, like $W^1, W^2, W^3, W^4 \ldots W^N$, $W^1+W^2+W^3+W^4+\ldots+W^N=1$

```
WHEN (space_reaches_watermark OR every_x_months):
    FOR EACH file IN DISK:
                           target_storage_class    =
decide_migration(file, file_attributes, ...)
        IF target_storage_class:
            MIGRATE_TO(file, target_storage_class)
```



Conclusion

- The exploratory data analysis provides useful patterns for data training
- The accuracy of prediction is up to 91.81%
- The policy engine is designed to optimize the data storage based on the predicted data popularity
- Next steps
    - Policy engine will be tested and integrated into the current storage
    - Testing model for degradation of accuracy over time
    - XGBoost hyperparameter optimization, allows more customizability for the data
    - Training more data with new labels, like 1 month hot, 1-6 month warm, 6+ month cold, etc
    - Test for other possible features that can be helpful to improve the model further

Thank you!

Backup

20 Clusters

Brookhaven National Laboratory

Prediction model and results

Model training: 60K

Accuracy: 0.9168333333333333

Classification Report:

		precision	recall	f1-score	support
Cold	0	0.93	0.91	0.92	4014
Warm	1	0.88	0.89	0.89	3963
Hot	2	0.94	0.95	0.94	4023
accuracy				0.92	12000
macro avg		0.92	0.92	0.92	12000
weighted avg		0.92	0.92	0.92	12000

Confusion Matrix:

$$\begin{array}{cccc} [[3660 & 306 & 48] & 0 \\ [\ 244 & 3525 & 194] & 1 \\ [\ 52 & 154 & 3817]] & 2 \\ 0 & 1 & 2 & \end{array}$$

Prediction model and results（cont.)

Top 300,000 access count

Accuracy: 0.9070202901840102

Classification Report:

		precision	recall	f1-score	support
Cold	0	0.88	0.87	0.88	20549
Warm	1	0.99	0.90	0.94	244456
Hot	2	0.60	0.99	0.75	34598
Accuracy				0.91	299603
macro avg		0.82	0.92	0.86	299603
weighted avg		0.94	0.91	0.91	299603

	Counts	Percentage
1	244456	0.815933
2	34598	0.115479
0	20549	0.068587

Confusion Matrix:

```
[[ 17920   2585     44]
 [  2347 219523  22586]
 [    54    241  34303]]
```

Prediction model and results（cont.)

Random 1,500,000

Accuracy: 0.9040399862300489

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.91	0.93	842949
1	0.90	0.89	0.89	630126
2	0.38	0.95	0.54	25841
accuracy			0.90	1498916
macro avg	0.74	0.92	0.79	1498916
weighted avg	0.92	0.90	0.91	1498916

	Counts	Percentage
0	842949	0.562372
1	630126	0.420388
2	25841	0.017240

Confusion Matrix:

```
[[770981  60067  11901]
 [ 41767 559450  28909]
 [   249    943  24649]]
```

Prediction model and results（cont.)

Random 1,500,000(Even # of each type)

Accuracy: 0.90868

Classification Report:

	precision	recall	f1-score	support
0	0.94	0.96	0.95	500000
1	0.84	0.93	0.88	500000
2	0.96	0.84	0.89	500000
accuracy			0.91	1500000
macro avg	0.91	0.91	0.91	1500000
weighted avg	0.91	0.91	0.91	1500000

Confusion Matrix:

[[481229 14049 4722]
[23935 464196 11869]
[8663 73742 417595]]

Prediction model and results

Total 6 months data: 23M

Accuracy: 0.9181894871283923

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.92	0.95	18371582
1	0.73	0.91	0.81	2699937
2	0.69	0.91	0.79	2022799
accuracy			0.92	23094318
macro avg	0.80	0.91	0.85	23094318
weighted avg	0.93	0.92	0.92	23094318

Confusion Matrix:

```
[[16903744   787431   680407]
 [   92564  2454124   153249]
 [   48712   126995  1847092]]
```