ベアメタルで実現する Spark＆Trino on K8sなデータ基盤

株式会社マイクロアド
永富 安和 (𝕏 @yassan168)

事業紹介（データプラットフォーム事業）

MicroAd
Redesigning the Future Life
#trinodb

DSP UNIVERSE Ads

広告を出したい「広告主」向け

広告主/代理店
- 広告代理店
- 広告主
- ：

広告
- 広告 A
- 広告 B
- ：

広告出稿料

リアルタイムで取引
(RTB)

AD ×

広告表示

SSP MicroAd COMPASS

広告を出して欲しい「Webメディア」向け

ユーザー
- ユーザーA
- ユーザーB
- ：

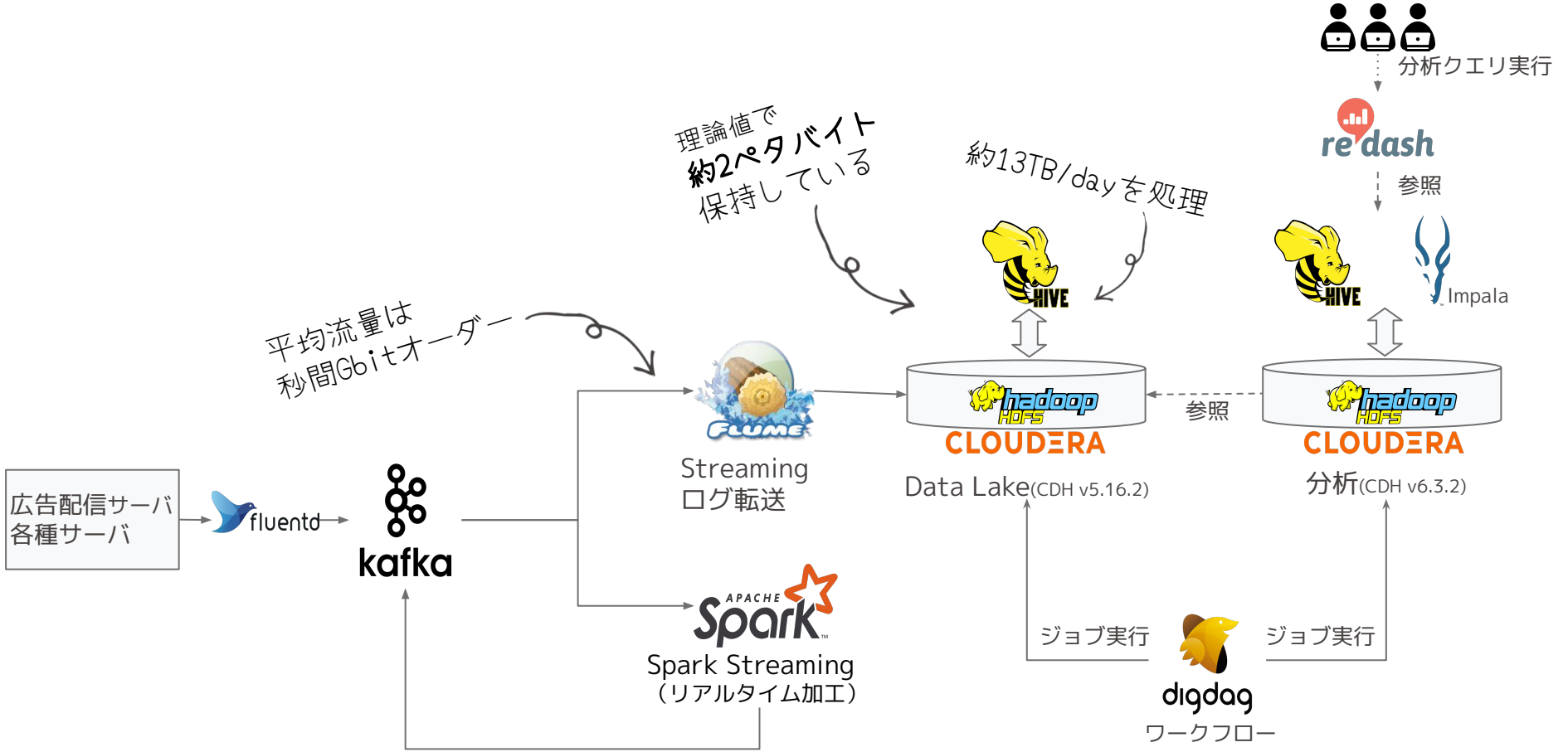
メディア
- ニュース
- グルメ
- ：

データ紐づけ

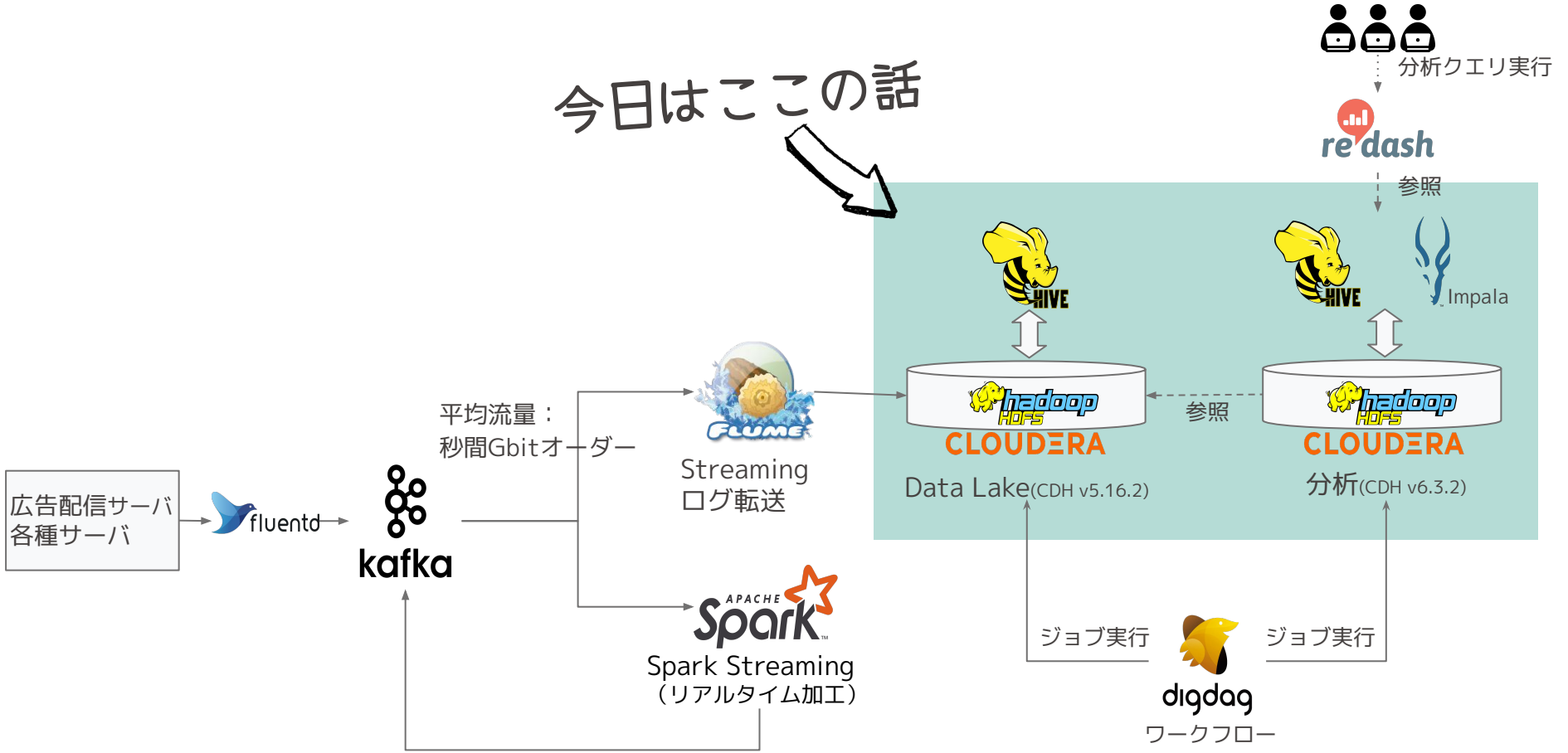
- 位置情報 提携企業DB
- EC購買 提携企業DB
- Web行動
- リアル購買 提携企業DB

Data Management Platform UNIVERSE

現行のデータ基盤の概要



現行のデータ基盤の概要



現行のデータ基盤の課題

1. **CDH無償版の提供が終了しているので継続して利用出来ない**
   - 有償の後継版Cloudera CDPも検討したが費用面がクリア出来ず見送り
     （Google Cloudなども検討したが、費用や技術課題がクリア出来ず見送り。5年償却で見るとクラウドは高い。）
2. **ComputeとStorageを分離してNode配置出来ないので**
   サーバスペックが過剰になりがち
   - YARNのNode ManegerとHDFSは分離して配置出来ない
   - ComputeスケールさせたいだけなのにStorageもスケールするので非効率
3. **Impalaの統計情報の運用が非常に煩雑かつ有効に利用出来ない**
   - 大規模テーブルの場合、ほぼ使えない
   - 統計情報が利用できないので効率の悪いクエリになりがちでImpalaを活かしきれない
4. **ETL/ELT処理で利用しているMapReduceベースのHiveが遅い**
   - 本来はMapReduceではなく、Tez・LLAPを使うべきだがCDHが古くて利用できない
5. **テーブル構造が複雑なので、SQLベースでETL/ELT処理するのが辛い**
   - 複雑なクエリになりがちで、改修に難易度が高く手間がかかる

新しいデータ基盤に求める事

1. ComputeとStorageを分離したい

2. ETL/ELT処理は、SQLベースではなく、Programmableに処理したい

3. SQLエンジンは大規模なテーブルでも統計情報を更新・有効活用が出来ること

4. Hiveテーブルの様にオンラインで柔軟なスキーマ進化が可能であること

新しいデータ基盤に求める事

1. **ComputeとStorageを分離したい**

   😆 HiveテーブルからIcebergテーブルに変更し、HDFSからS3互換のアプライアンスに置き換えることでYARN・Zookeeperに依存しなくなり分離が可能になった（構成要素も減ったので構築も楽になった）

2. **ETL/ELT処理は、SQLベースではなく、Programmableに処理したい**

   😆 Sparkを使ってスクリプトベースに処理することで、複雑なSQLでの処理が不要になった

3. **SQLエンジンは大規模なテーブルでも統計情報を更新・有効活用が出来ること**

   😆 Trino＆Icebergを使うことで、Hive・Impalaに依存せずに、柔軟に統計情報の更新・利用する

4. **Hiveテーブルの様にオンラインで柔軟なスキーマ進化が可能であること**

   😆 Iceberg特有のスキーマ進化（orパーティション進化）により、以前より柔軟な運用が可能になる

新しいデータ基盤の概要

MicroAd
Redesigning the Future Life
#trinodb

Storage

S3互換ストレージ

Compute

trino

Coordinator　Worker　・・・　Worker

RKE2

Node 1　Node 2　・・・　Node N

APACHE Spark

Spark App
Spark App
Spark App

Driver　Executor　・・・　Executor

RKE2

Node 1　Node 2　・・・　Node N

Downstream User Cluster を管理

Rancher Server

Pod

新しいデータ基盤の概要

MicroAd
Redesigning the Future Life
#trinodb

Storage

S3互換ストレージ

今日はここにフォーカス

Compute

trino

Coordinator　Worker　・・・　Worker

RKE2

Node 1　Node 2　・・・　Node N

APACHE Spark

Spark App
Spark App
Spark App

Driver　Executor　・・・　Executor

RKE2

Node 1　Node 2　・・・　Node N

Downstream User Cluster を管理

Rancher Server

Pod

どうやって切り替えていくか？

予算やデータセンタの設備の問題、技術的な課題などなど、いろんな理由から、
分析用クラスタ→Data Lake用クラスタの順に切り替えていきます。

詳しい話を始めると枠にまったく収まらないので、
マイクロアドの技術ブログなどで発信して行きます！

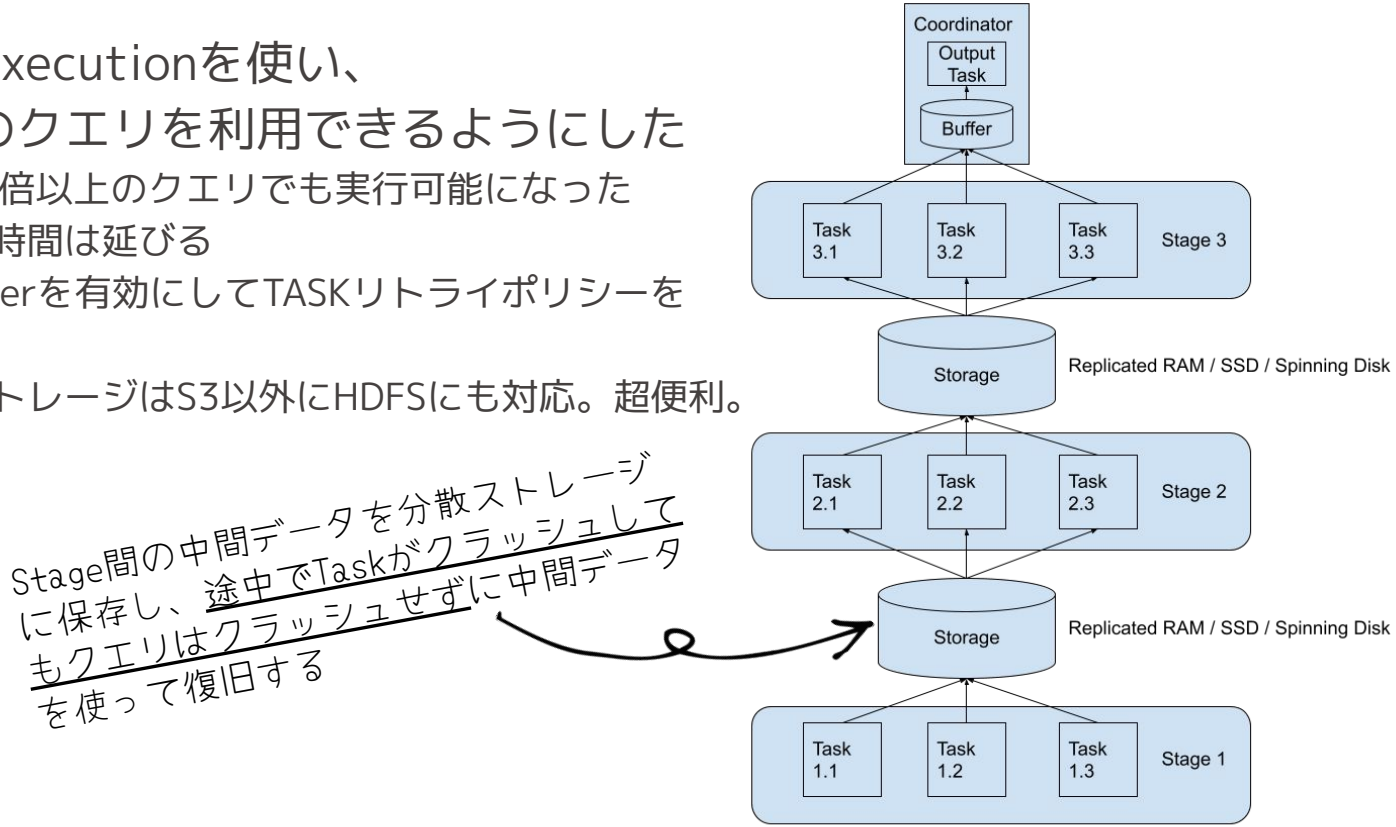
もしくは10/12のOTFSG Tokyo Meetup #1
https://otfsg-tokyo.connpass.com/event/296440/ に
参加するので、そこで捕まえてください。

アドホック分析用としてのTrinoで工夫したこと

- Kubernetes（RKE2）を使うことでクラスタの構築やアップデートを楽にした
  - Trino自体の構成がCoordinator・Workerと構成がシンプルでPersistentVolume (PV)が不要。その為、マニフェストをシンプルに保てるのでK8sでの運用はさほど辛くない。
  - RKE2のsystem-upgrade-controllerがあるので設定書いてapplyするとローリングアップグレードしてくれるので便利 https://docs.rke2.io/upgrade/automated_upgrade/
    （もしくはRancher Web UIからポチーがもっと簡単）
- Rancherを使ってK8sクラスタ管理することで管理コストを下げ利便性を向上
- Helm ChartにはTrino公式のものよりもこなれている github.com/valeriano-manassero/helm-charts を使用した
  - JVMのヒープサイズの指定を-Xmx/-Xmsではなく-XX:MaxRAMPercentage/-XX:InitialRAMPercentageを使って使用可能なメモリーに対する割合で指定するように変更
  - Affinityを利用して、CoordinatorとWorkerポッドの同居を禁止し、Workerポッドはなるべく同一Nodeに2個以上配置しないようにする

アドホック分析用としてのTrinoで工夫したこと

- Fault-tolerant executionを使い、
  搭載メモリ以上のクエリを利用できるようにした
  - 搭載メモリの10倍以上のクエリでも実行可能になった
    ※ただし、実行時間は延びる
  - Exhange managerを有効にしてTASKリトライポリシーを使用しました。
  - 中間データ用ストレージはS3以外にHDFSにも対応。超便利。



アドホック分析用としてのTrinoで工夫したこと

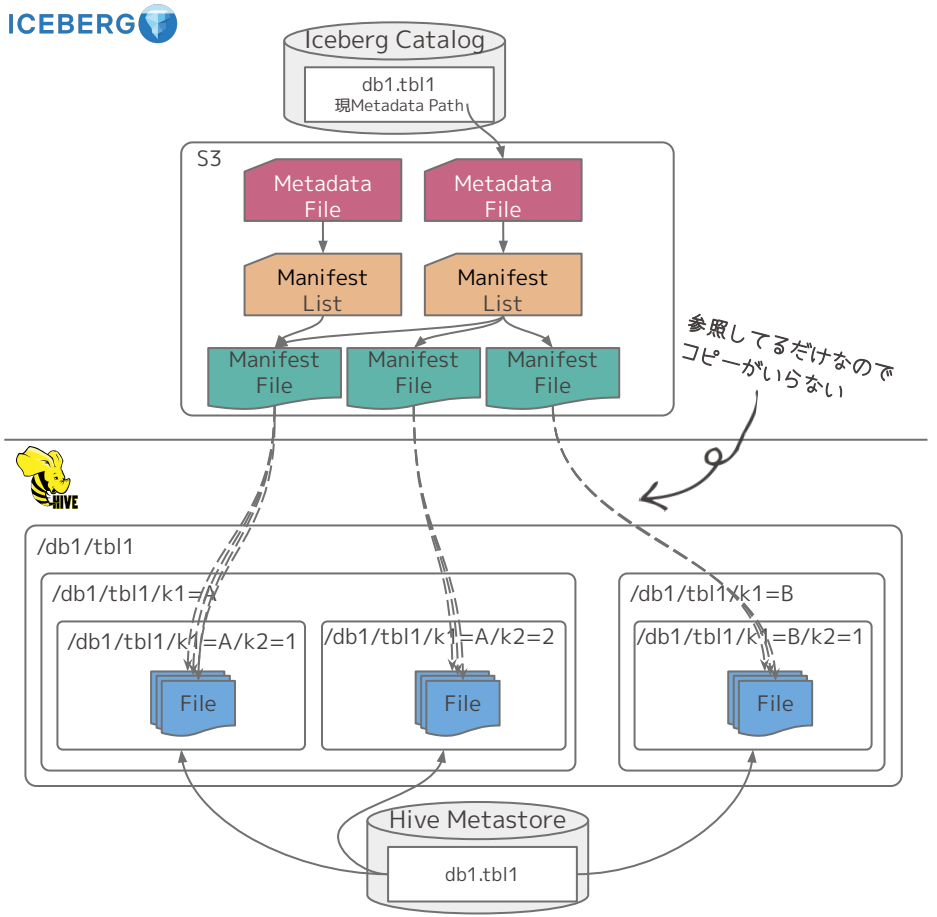
- Spill-to-diskを使って、OOMを起きにくくするようにした
  - Helm Chartを改修して、Coordinator・WorkerポッドにemptyDirボリュームをマウントすることで、OOMで落ちても復帰時に再利用出来るようにした
- Icebergテーブルで利用するカタログにはRESTカタログを用うことで、TrinoやSparkなどからIcebergテーブルを利用しやすくした
- Hive→Icebergテーブル移管の際は、IcebergのSparkのadd_filesプロシージャを使うことで、Icebergテーブル側に過去分のデータをコピーを不要にした
  - HiveテーブルはHDFS上にあるので、add_filesプロシージャで出来たIcebergテーブルのデータは、S3とHDFSの両方を参照することになるので、カタログ（REST）のio-implプロパティにorg.apache.iceberg.io.ResolvingFileIOを利用することで両方に対応した

補足：add_filesプロシージャって？

移行元のデータをIcebergテーブルにコピーせずに参照出来るようにするIcebergのSparkのプロシージャ。

数万パーティションある様なテーブルの場合、一度にコピーするには時間がかかるが、これなら移行元の更新を止めずに移行が可能。

パーティション単位で実行が可能なので、移行＆検証が終わるまでの期間はadd_fileを使い追加分を更新し、準備が終わったらIcebergテーブルにデータ書いていけば良いので移行作業の効率が良い。

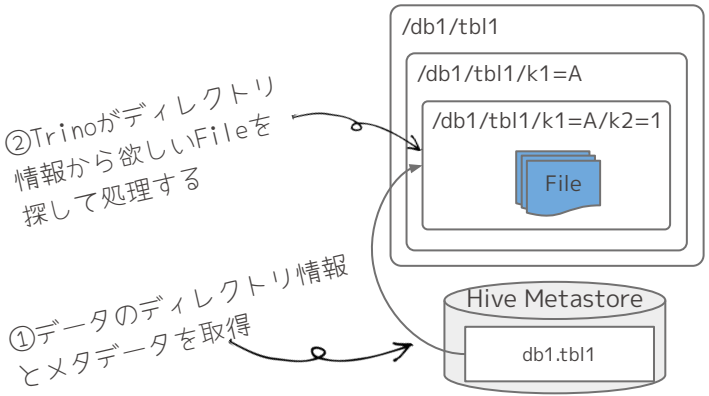


補足：TrinoはどうやってHiveやIcebergテーブルを参照するの？

Hiveテーブルの場合

thrift経由でHive Metastoreに対して、メタデータとデータの格納先ディレクトリを取得するだけで、
実際のデータ処理はTrino側で実施。

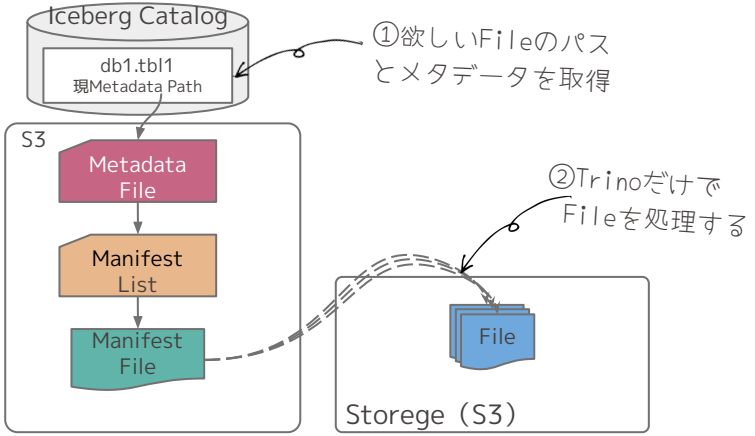
その為、YARNなどは一切関与しない。



Icebergテーブルの場合

カタログから欲しいFileの全パスとメタデータを取得し（取得方法はカタログ実装によりけり）、その情報を元にTrino側で処理を実施。

Icebergは欲しいFileパスが①の段階で確認出来る
（HiveはHDFSディレクトリまでなので②のタイミングで探す必要がある）



現状、困っていること

1. Icebergのadd_filesプロシージャが
   TrinoのIcebergコネクターから利用できない
   - Add `add_files` procedure in Iceberg connector · Issue #11744 · trinodb/trino

2. Icebergのadd_filesプロシージャを使ったテーブルにて、
   参照元のHiveテーブルにtimestamp型があった場合、
   そのままでは以下のエラーが出て参照出来ない
   - エラー文：Query 20231002_061128_00067_tqdd2 failed: Unsupported Trino column type (timestamp(6) with time zone) for Parquet column ([update_time] optional **int96** update_time = 8)
   - Trino Iceberg not honoring existing timestamp column type name of the table created outside Trino (e.g. Spark) stored in HMS · Issue #11442
   - Hive/Impalaで利用しているTIMESTAMP型はINT96でParquetファイルで書き込みしているがTrinoで利用しているParquetライブラリは新しくINT96に対応していない事が影響

今後の予定

1. 9月末にリリースのあったTrino Gateway trinodb/trino-gateway を利用する
   - Trinoクラスタを2系統用意してTrino Gateway経由で利用する
   - Trinoの設定反映やアップグレードの際に片系ずつ実施する事が可能になるのでサービスのダウンタイムをなくすことが出来る
   - TrinoのCoordinatorはHA構成が取れない（補足を参照）ので、Trinoサービスとしての可用性向上の目的の意味もある

2. 利用状況に合わせたResource groupsとSession property managerの設計
   - クエリ実行時間の制限（連続XX時間まで）
   - ユーザに合わせたクエリ種別の制限
     - XXXユーザは特定のテーブルにはSELECTのみに限定
   - 分析クラスタ内でのバッチへのリソース割当てを最優先にする

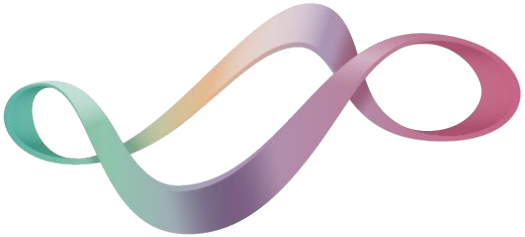
3. Icebergの統計情報を使ったパフォーマンス改善と運用整備

以下で情報発信をしています！

X（旧Twitter）
@microad_dev

技術ブログ
developers.microad.co.jp

We Are Hiring!!



MicroAd
Redesigning the Future Life

オンプレ×GCPな大規模データプラットフォームの開発・運用を
一緒に挑戦してみたい人を募集しています！

https://recruit.microad.co.jp/

公式Xアカウント **@microad_dev** もよろしくお願いします。

補足

- Fault Tolerant Executionに関する情報
  - wikiにある公式ドキュメントには無い詳細な説明
    https://github.com/trinodb/trino/wiki/Fault-Tolerant-Execution
  - Trino | Using Trino as a batch processing engine
    https://trino.io/blog/2022/06/24/trino-meetup-extract-trino-load.html
  - Trino | Project Tardigrade delivers ETL at Trino speeds to early users
    https://trino.io/blog/2022/05/05/tardigrade-launch.html
- TrinoのHAに関連する情報
  - Can you set up Trino in HA mode? - Trino - Starburst forum
    https://www.starburst.io/community/forum/t/can-you-set-up-trino-in-ha-mode/31
  - High Availability · Issue #391 · trinodb/trino
    https://github.com/trinodb/trino/issues/391
- Icebergについて深く知る事が出来る良い記事
  - Apache Iceberg: An Architectural Look Under the Covers | Dremio
    https://www.dremio.com/resources/guides/apache-iceberg-an-architectural-look-under-the-covers/

- Hive→Iceberg移管に関して参考になるブログ記事
  - How to Migrate a Hive Table to an Iceberg Table | Dremio
    https://www.dremio.com/blog/how-to-migrate-a-hive-table-to-an-iceberg-table/
  - Migrating a Hive Table to an Iceberg Table Hands-on Tutorial | Dremio
    https://www.dremio.com/blog/migrating-a-hive-table-to-an-iceberg-table-hands-on-tutorial/
- 利用しているIcebergのREST Catalog実装
  - https://github.com/tabular-io/iceberg-rest-image
  - Iceberg's REST Catalog: A Spark Demo • Tabular
    https://tabular.io/blog/rest-catalog-docker/
- Icebergを知りたいならここから始めると参考になる記事
  - Apache Iceberg 101 - Your Guide to Learning Apache Iceberg Concepts and Practices | Dremio
    https://www.dremio.com/blog/apache-iceberg-101-your-guide-to-learning-apache-iceberg-concepts-and-practices/
  - Apache Iceberg FAQ | Dremio
    https://www.dremio.com/blog/apache-iceberg-faq/#h-what-is-a-data-lakehouse

- ParquetファイルのINT96関連情報
  - parquet-format/LogicalTypes.md
    https://github.com/xhochy/parquet-format/blob/cb4727767823ae201fd567f67825cc22834c20e9/LogicalTypes.md#int96-timestamps-also-called-impala_timestamp
  - Parquet: Support filter operations on int96 timestamps by thesquelched · Pull Request #2563 · apache/iceberg
    https://github.com/apache/iceberg/pull/2563
  - 'NOT_SUPPORTED: Unsupported Trino column type (date) for Parquet column ([today] optional int64 today (TIMESTAMP(MICROS,false))) · Issue #17733 · trinodb/trino
    https://github.com/trinodb/trino/issues/17733
- S3互換ストレージと言えばMinIO以外にもApache Ozoneもあるよ（宣伝）
  - S3互換のオブジェクトストレージ Apache Ozoneに関する情報（随時更新） - Qiita
    https://qiita.com/yassan168/items/1e3c000284ae6fc8448c

- RKE2
  - https://docs.rke2.io/
- Rancherを利用したモニタリング＆アラート
  - https://ranchermanager.docs.rancher.com/pages-for-subheaders/monitoring-and-alerting
- Rancher
  - https://www.rancher.com/
  - 日本のユーザコミュニティもあるのでよろしくです。
    - https://rancherjp.connpass.com/