HITACHI
Inspire the Next

オープンソースカンファレンス2019 .Enterprise

今さら聞けない！HadoopをSIで使いこなすコツ
～サポートエンジニアが考えるオンプレ対応のポイント～

2019/10/10

株式会社 日立製作所
サービスプラットフォーム事業本部 OSSソリューションセンタ

木下翔伍

はじめに

HITACHI
Inspire the Next

このセミナー

Hadoopをオンプレ環境で構築することになったSI案件を推進する上で、
躓きやすい点と対処の方法や考え方を解説します

対象の聴講者

Hadoopを使ったシステム開発に関わる（特にオンプレ環境の経験が少ない）人
Linuxの基本的な知識や操作経験はある人

セミナー資料

後日、OSCのWEBサイトで公開予定です

なぜオンプレの話か

HITACHI
Inspire the Next

トレンド

Hadoopを扱う案件では、**パブリッククラウド**ベンダ提供の(マネージド)サービスを活用して開発を進めたり、PoCや技術検証に取り組むケースが見られます

実際のところ

顧客要件によって「**本番はオンプレ**」となるケースもまだまだあります
すべてがクラウド(オンプレがゼロ)にはならないと考えています

ギャップ(躓きやすい点)

- クラウドに任せていたこと(特にインフラ設計)を自前で対応する場面に直面します
- WEB等での情報は豊富ですが、Hadoopに慣れた(使い込んでいる)人でないと必要な情報を検索して理解することが難しいです

このギャップを埋めるべくテクニカルサポートによく寄せられるお問い合わせからいくつかのポイントをピックアップしてサポートエンジニアがご紹介します

講演者プロフィール

HITACHI
Inspire the Next

木下翔伍 / Shogo Kinoshita

やっていること

ビッグデータ関連ソリューションの検討・開発

Hadoopとその関連OSSを活用した技術検証

例）
デジタル電力計1,000万台分のデータを活用して
電力設備にかかる負荷を推測するユースケースにおいて、
Sparkでデータを処理したときの性能を検証

検証結果を講演・記事執筆し、一部は書籍化

テクニカルサポートサービス

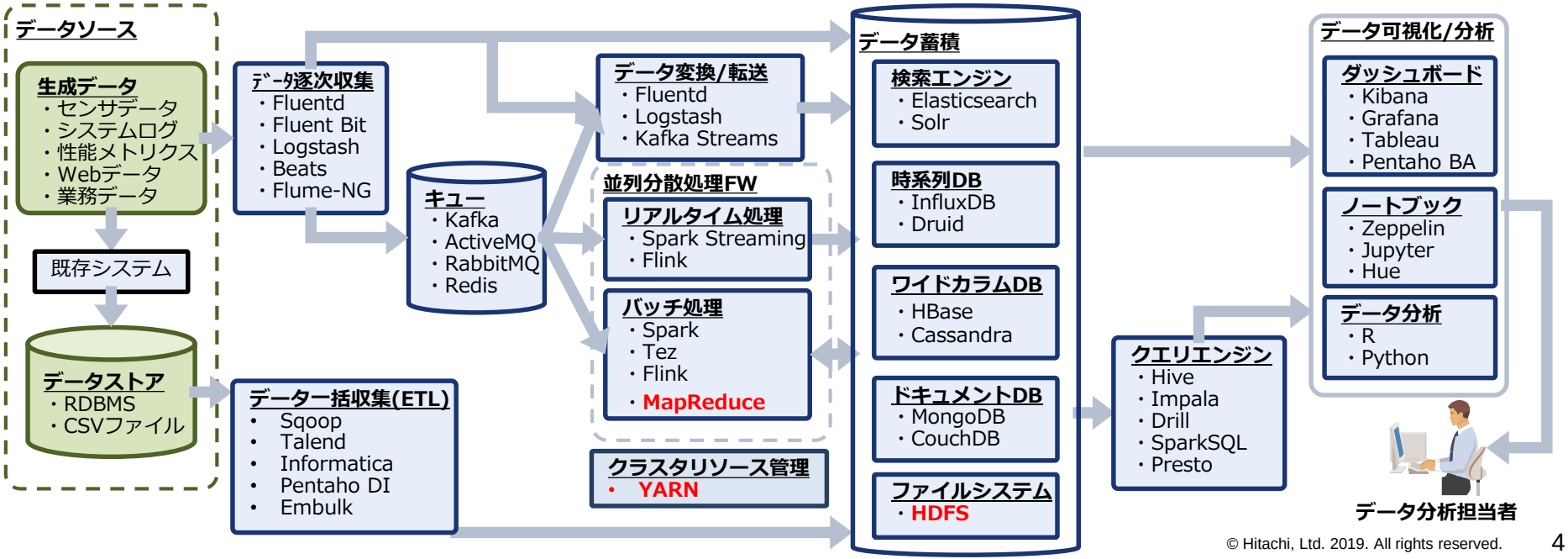
顧客から寄せられる技術的な問題に対し、必要に応じて
調査・検証を実施して、対処方法や回避策をご提案

Q. 1　Hadoopとは何ですか

HITACHI
Inspire the Next

A. 1

- 多種/多量のデータを蓄積し、高速に分散処理するための基盤となるOSSミドルウェアです
- 3つのコンポーネントで構成されます(HDFS:分散ファイルシステム,
  YARN:クラスタリソース管理, MapReduce:並列分散処理フレームワーク)
- Hadoopとその周辺で活用される多数のOSS(Hadoopエコシステムと呼ばれる)を
  組み合わせることで、データの収集から蓄積、分析、結果の可視化までをカバーしています

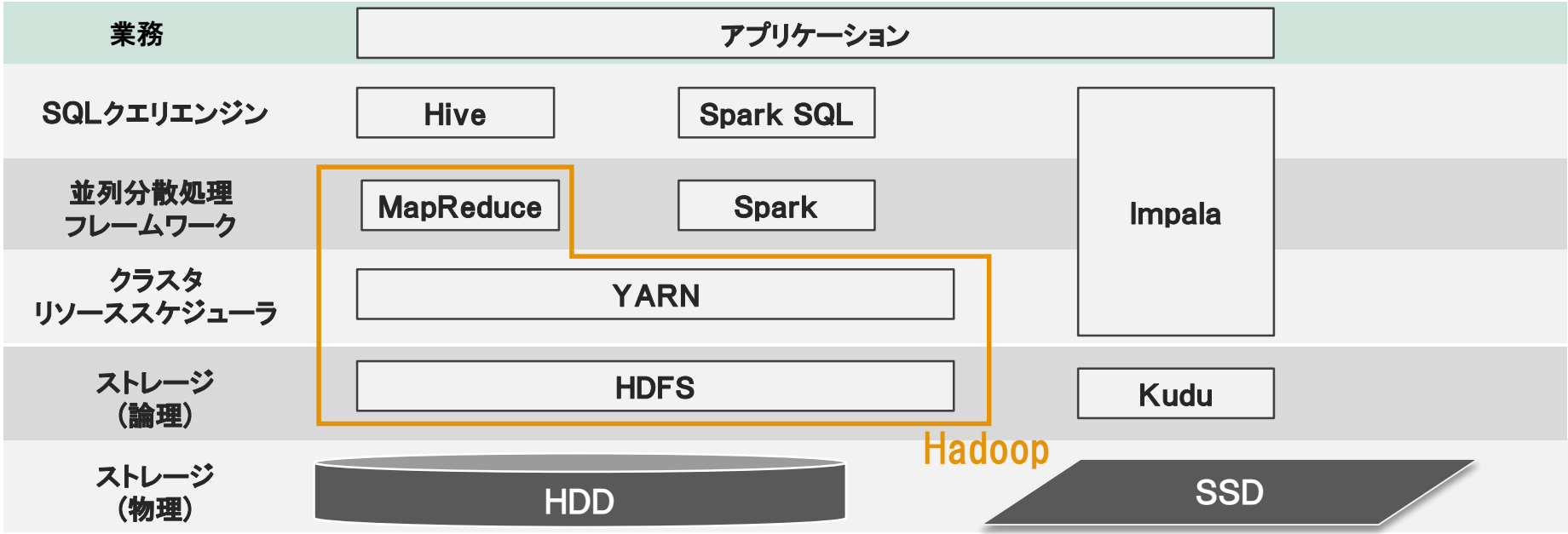


Q. 2 SQL on Hadoopとはどのようなものですか

HITACHI
Inspire the Next

A. 2

- 明確に定義されてはいませんが、構造化されたデータを対象とするHadoopの並列分散処理をSQLを用いて実現するための技術全般のことをさします
- Hive, Impala等がHadoopにおけるSQLクエリエンジンとしてよく知られており その周囲では次のようなソフトウェアが関わっています



Q. 3 ノード数はどのくらい必要ですか

HITACHI
Inspire the Next

A. 3

- 少なくともマスタノードは3ノード以上、ワーカノードは4ノード以上から検討します
- メタストアDB(テーブルのメタデータ管理)やクライアント用のホストは別途必要です
- リソースや性能が不足するようであればノードを追加します(スケールアウト)

マスタノードのノード数

HDFS NameNodeやYARN ResourceManagerを冗長化(HA)構成にするとき、
ノード数は奇数かつ3ノード以上であること
・HA構成の前提ソフトウェアであるZooKeeperを構築するとき、
3ノード以上の奇数ノードで構築されている必要があるため

ワーカノードのノード数

HDFSレプリケーション係数[dfs.replication](デフォルト値3)よりも大きいほうがよい

Q. 4　ホストに必要なリソースはどれくらいですか

HITACHI
Inspire the Next

A. 4

業務によりますが、見積りの考え方や検討の初期値になりそうな値を
リソースごとに挙げます

CPUコア数（マスタノード）	ホストが搭載するディスク数 + 2 [個] 以上
CPUコア数（ワーカノード）	DataNodeに割り当てるディスク数 + 2 [個] 以上

・ホストが搭載するディスク数と同じコア数をディスクI/Oに用意します
・OSやHadoopの各種デーモンに使わせるコアを2個くらい用意します

Q. 4　ホストに必要なリソースはどれくらいですか

HITACHI
Inspire the Next

ディスク数（マスタノード）	6 [本] 程度以上
ディスク数（ワーカノード）	**ホストに搭載できるだけめいっぱい**（少なくとも5本）

・具体的には用途（ホストに割り当てるロール）を考慮して検討します
・一般にディスク数は多いほうがHDSFのI/O性能の観点で有利です

関連：Q. 8

ネットワークインタフェース	10 [Gbps] に対応したネットワークインタフェース（と回線）を導入

・特にワーカノード間では分散ジョブの結果集約等で通信が
大量に発生するので、対応できるように十分な帯域を確保します

Q. 4　ホストに必要なリソースはどれくらいですか

HITACHI
Inspire the Next

メモリ容量（マスタノード）

（HDFSブロック数 / 100万）+ 8 [GB] 以上

HDFSでのブロックとは、格納されるファイルをある容量(CDHのデフォルトでは128MB)ごとに分割したデータのことです

・HDFS NameNodeのメモリ領域に格納されるファイルシステムイメージ（名前空間;メタデータ)のサイズよりも大きな容量を用意します
・OSやHadoopの各種デーモンに使わせる容量を用意します

HDFSメタデータ量の見積り
　1ファイルごとに（1 + ファイルのブロック数）× 150 [Byte]　が目安

実際には
「HDFSが100万ブロックを持つとき、NameNodeはメモリ1GBを使用する」と見なすことが多い
(※)ブロック数には複製されたものを含まない

補足　NameNodeが必要とするメモリ量

HITACHI
Inspire the Next

1ファイルあたりのメモリ量の目安は （1 + ファイルのブロック数）× 150 [Byte] なので
HDFSのブロックサイズが128MBの場合・・・

・128MBのファイルを1個格納するとき

（1 + 1[ブロック]）× 150[byte] × 1[ファイル] = 約300[byte]

・1MBのファイルを128個格納するとき

（1 + 1[ブロック]）× 150[byte] × 128[ファイル] = 約38,400[byte] (=約38.4 [KB])

・192MBのファイルを2個格納するとき

（1 + 2[ブロック]）× 150[byte] × 2[ファイル] = 約900[byte]

Q．４　ホストに必要なリソースはどれくらいですか

HITACHI
Inspire the Next

メモリ容量（ワーカノード）

PoCなどで事前に検証を実施して見積もる

・並列分散処理の対象となるデータと処理途中に発生する一時データを
すべてメモリに載せられるように容量を用意します

ただし処理対象のデータや処理内容によってメモリ消費量は
大きく異なることから、データ処理に必要となるメモリ量を事前に机上で
見積もることは困難です

PoC等の検証において、実際の業務で使用するデータのサブセットに対して
実際の業務の処理を実行し、処理対象のデータと処理中に発生する
一時データの総和を見積もることをおすすめします

Q. 4　ホストに必要なリソースはどれくらいですか

HITACHI
Inspire the Next

ディスク容量（HDFS）	（HDFSに格納するデータ量）× 3 ÷ 0.8

・HDFSのデータ格納領域として必要なディスク容量を用意します
（上記計算式はワーカノード全体で求められるディスク容量の目安）

レプリケーション係数はデフォルトが3なので、
格納するデータ量の約3倍のディスク容量が実際には必要です

さらに一時ファイル出力やバッファを考慮した余裕が必要なので、
ここでは約20%の余裕を持たせます

Q. 5　クラスタのネットワークを構築するときの注意点はありますか

HITACHI
Inspire the Next

A. 5　代表的な注意点を3点挙げます

Hadoop用と一般業務用にネットワークを分離します

・特にワーカノード間のトラフィックが、既存の他システムに影響を与えることを排除します

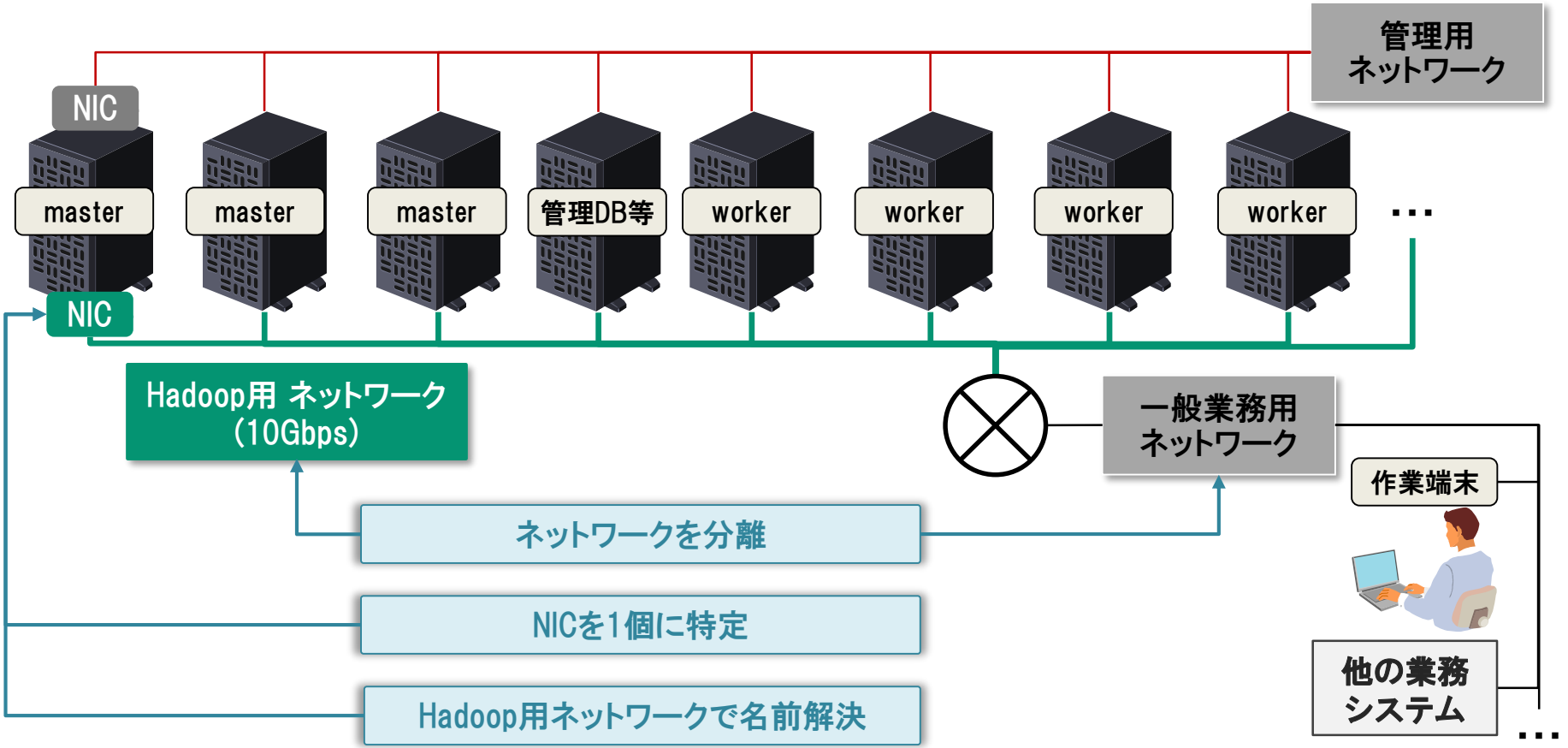
Hadoopで使うNIC（ネットワークインタフェース）は1個に特定します

・通信先によってネットワークを使い分ける機能はまだサポートされていません

名前解決はHadoop用ネットワークに接続されたインタフェースでします

次のスライドで、ネットワーク構成の具体例を図解します

Q. 5　クラスタのネットワークを構築するときの注意点はありますか



Q. 6　クラスタのホスト名に何か制約はありますか

HITACHI
Inspire the Next

A. 6

・hostnameコマンドで返る値が次の点を満たすようにしてください

・クラスタ内で一意な値を設定します

　この「クラスタ」には、Backup and Disaster Recovery機能を使うときのバックアップ環境と管理DBのホストも含みます

・英小文字と数字の組み合わせのみで設定します

・Sentry(Active Directory)を導入する可能性があるとき、あらかじめ15文字以内に設定します

Q. 7　DNSではなくhostsファイル編集で名前解決してよいですか

HITACHI
Inspire the Next

A. 7

・可能ですが、次の点に注意して設定します

・Hadoop用ネットワークに接続されたNIC（ネットワークインタフェース）を使うこと

・エイリアスにFQDNを使わないこと

　ホスト名がFQDNで設定されることや、エイリアスを使用しないことは問題ありません

Q. 8　ホスト（OS）上でのディスク（ボリューム）はどのように設定しますか

HITACHI
Inspire the Next

A. 8　ノードや領域の用途によって異なるので、構成の一例をそれぞれ挙げます

OS領域 （ノード共通）	ディスク2本（RAID1：ミラーリング）で構築したボリューム

・クラスタを構成するすべてのホストが対象です
・OSがインストールされる領域のため冗長化されていることが望ましいです

Q. 8 ホスト(OS)上でのディスク(ボリューム)はどのように設定しますか

HITACHI
Inspire the Next

マスタノードに特有のボリューム

ディスク2本(RAID1:ミラーリング)で構築したボリューム1個

・NameNodeデータディレクトリとして運用します

ディスク1本で構築したボリューム1個

・JournalNode editsディレクトリとして運用します

ディスク1本で構築したボリューム1個

・ZooKeeperデータディレクトリ及びトランザクションログディレクトリとして運用します

Q. 8 ホスト(OS)上でのディスク(ボリューム)はどのように設定しますか

HITACHI
Inspire the Next

ワーカノードに特有のボリューム	OSインストール領域に用いていないディスクをすべてJBOD(もしくはディスク1本でのRAID0)で構築したボリューム

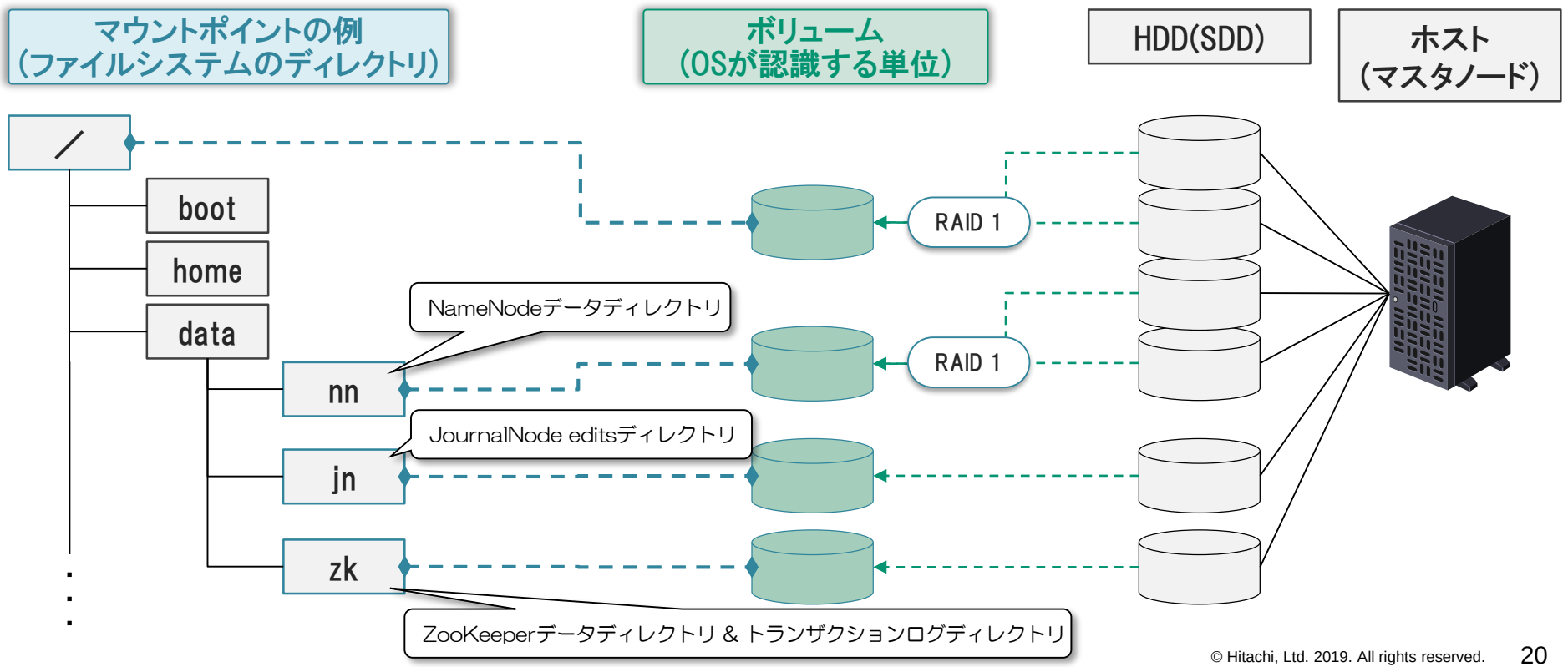
- DataNodeデータディレクトリ及びNodeManagerローカルディレクトリとして運用します

次のスライドで、ボリュームのマウント先の一例をノードごとに図解します

Q. 8 ホスト(OS)上でのディスク(ボリューム)はどのように設定しますか

HITACHI
Inspire the Next

ボリュームとディレクトリ設計の例 (マスタノード)



Q. 8　ホスト(OS)上でのディスク(ボリューム)はどのように設定しますか

HITACHI
Inspire the Next

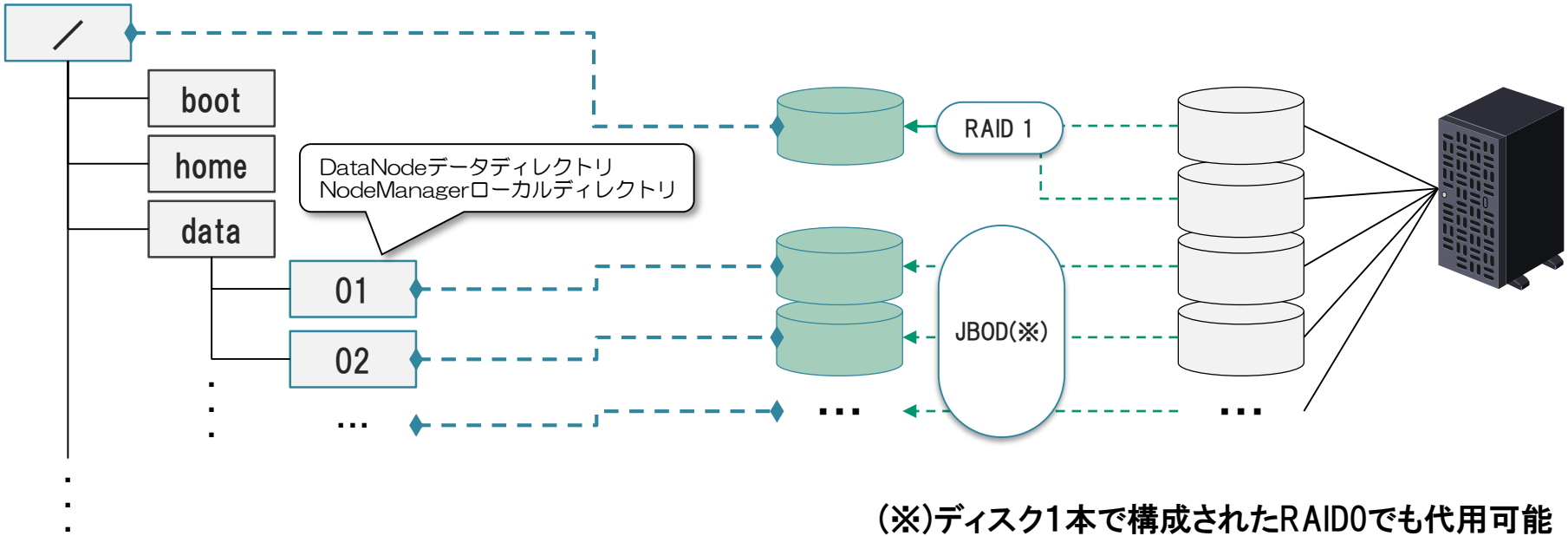
ボリュームとディレクトリ設計の例 (ワーカノード)

マウントポイントの例
(ファイルシステムのディレクトリ)

ボリューム
(OSが認識する単位)

HDD(SDD)

ホスト
(ワーカノード)



(※)ディスク1本で構成されたRAID0でも代用可能

Q. 9 Impalaログのローテートが設定と異なる挙動なのですが

HITACHI
Inspire the Next

A. 9

・Impalaが頻繁にクラッシュ(自動再起動)する環境では、
ファイルサイズが上限に満たないログが生成されたり、バックアップ数を超えて
ログファイルが保持されることがあり得ます

この事象の原因

ImpalaはHDFSやYARNとはログローテートの処理ステップが異なるためです

次のスライドで、処理ステップを図解します

そのまま放置した場合、次のような事象に発展する可能性があります

・バックアップ数を超えたファイルが保持されることで、ディスク容量が想定以上に消費される
・想定よりも短い期間のログしか保持できていない

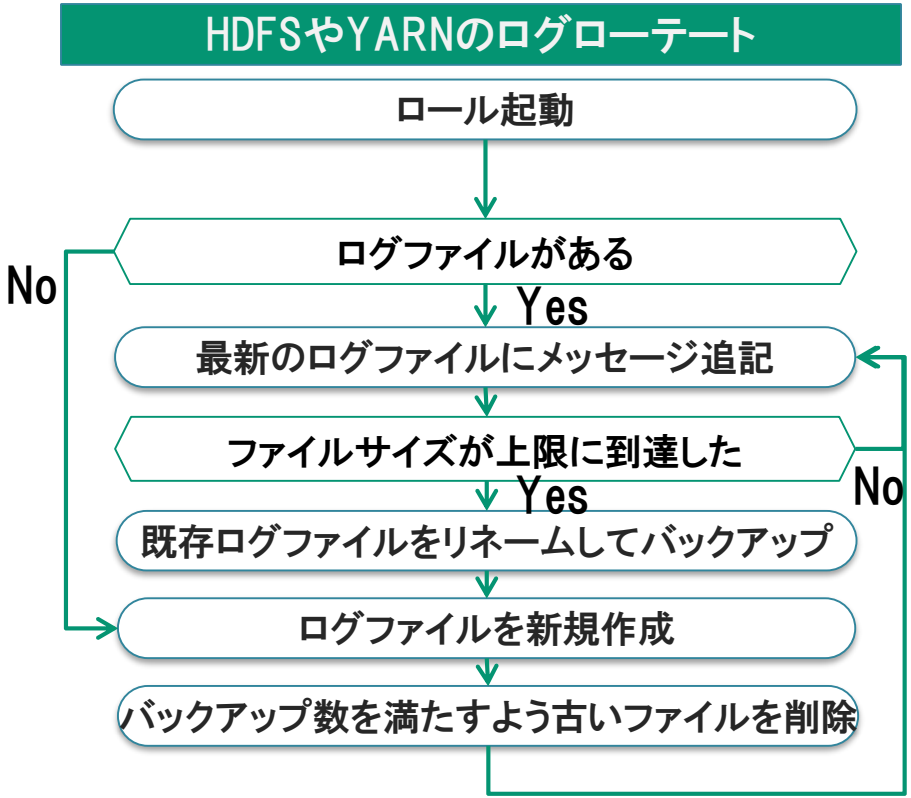
対応案

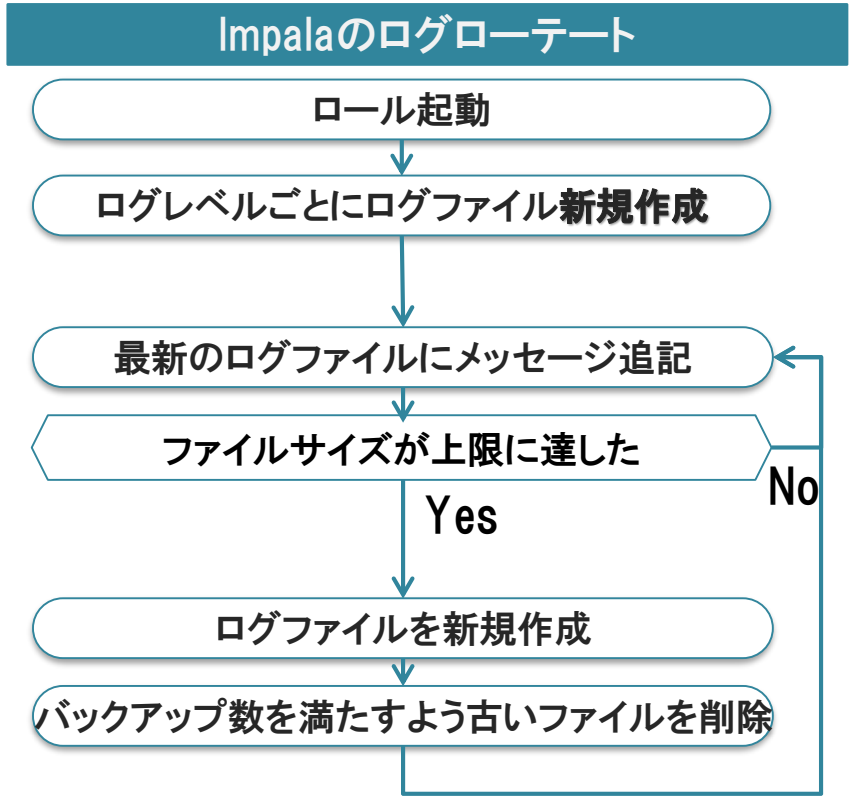
ログのバックアップ数を増やし、ファイルサイズを小さくすることを検討します

Q. 9 Impalaログのローテートが設定と異なる挙動なのですが

HITACHI
Inspire the Next

起動直後にImpalaはログファイルを新規に作成するが古いログの削除はしない





他社商品名、商標等の引用に関する表示

HITACHI
Inspire the Next

- Hadoopは、Apache Software Foundationの米国およびその他の国における登録商標または商標です。
- Clouderaは、Cloudera, Inc.の米国および他国の管轄権における商標もしくは登録商標です。
- Linuxは、Linus Torvalds氏の日本およびその他の国における登録商標または商標です。
- Red Hat, and Red Hat Enterprise Linux are registered trademarks of Red Hat, Inc. in the United States and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.
- Active Directoryは、米国Microsoft Corporationの米国およびその他の国における登録商標または商標です。
- その他記載の会社名、製品名などは、それぞれの会社の商標もしくは登録商標です。

HITACHI

Inspire the Next