OpenStack Juno
Sahara(Hadoop-as-a-Service)の概要

日本ヒューレット・パッカード株式会社
テクノロジーコンサルティング事業統括本部
オープンソース部

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

自己紹介: 大矢 俊夫（おおやとしお）

- 所属
  - 日本ヒューレット・パッカード株式会社
    テクノロジーコンサルティング事業統括本部 サービス統括本部 オープンソース部
- 職務領域
  - OSSおよびJava製品関連に全般に関する技術支援、コンサルティング
    - OpenStack、Hadoop、Java Application Server、OSS DB 等々
- 経歴
  - HP-UX OEM向けサポート
  - 非接触ICカード向けサーバミドルウェア設計・開発
  - 2008年から現職

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Agenda

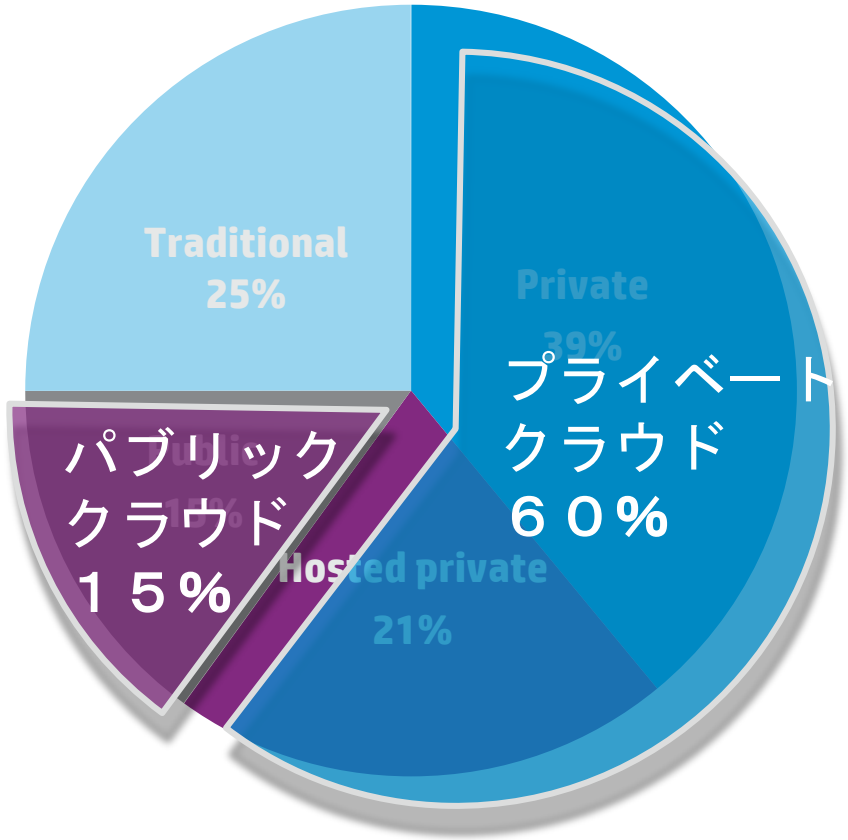
- OpenStack概要
- Hadoop-as-a-Serviceとは
- Sahara概要
- デモ
  - Pig を利用したビッグデータ処理
- 参考情報

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

OpenStack概要

企業ITはハイブリッドクラウドに向かう

2016年の企業IT配備モデル（予測） [1]



“..2017年末には、大企業の半数はハイブリッドクラウドを活用しているだろう”– Gartner [2]

[1] Source: Coleman Parkes Research by HP, May 2013

[2] “Gartner Press Release: ”Gartner says Cloud Computing Will Become The Bulk of New IT Spend By 2016” , Goa, India, October 24, 2013. http://www.gartner.com/newsroom/id/2613015

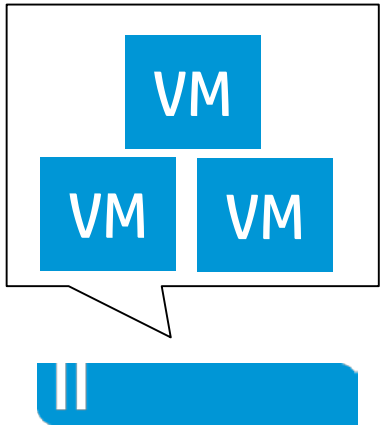
© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

IT基盤が「クラウド化」へと進化する背景

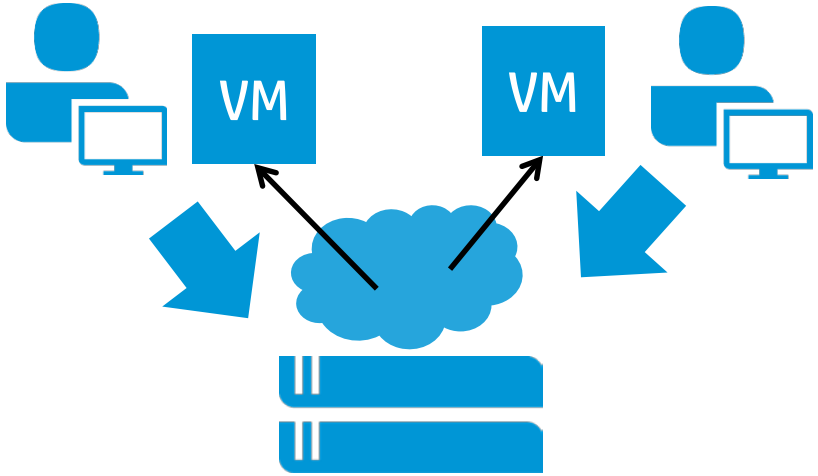
「仮想化」への進化、「クラウド化」への進化

仮想化によるIT基盤統合 → IT基盤のクラウドサービス化

サイロ型IT基盤



仮想化技術を活用した
IT基盤統合と標準化



IT基盤のクラウドサービス化

得られる効果

- リソース稼働率の向上
- 運用作業の標準化
- システムコストの最適化

- ITリソースを、サービスメニュー化して迅速に提供
- セルフポータルの提供による管理業務の自動化

「クラウド化」したIT基盤を構築できるソフトウェアのニーズが拡大

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

IT基盤の「クラウド化」とは

- ITリソースの「サービス化」＋「標準化」＋「自動化」
  - ✓ 実装手段として「仮想化」技術を利用することが多いが「仮想化」は必須ではない

サービス化

- 利用者はIT基盤の内部構造を意識しない
- 使いたいときに使いたい分を利用する

標準化

- 次のような条件を共通メニューとして揃える
  - ✓マシンリソース要件（OS、CPU、メモリ等）
  - ✓利用条件(SLA、セキュリティ等)
  - ✓申請方法、運用管理等のプロセス

自動化

- 利用申請やリソース払い出しなどの管理タスクをポータルやAPIで自動化

IT基盤の利用者のメリット

- 要求に応じたスペックの仮想サーバやストレージをすぐに利用できる

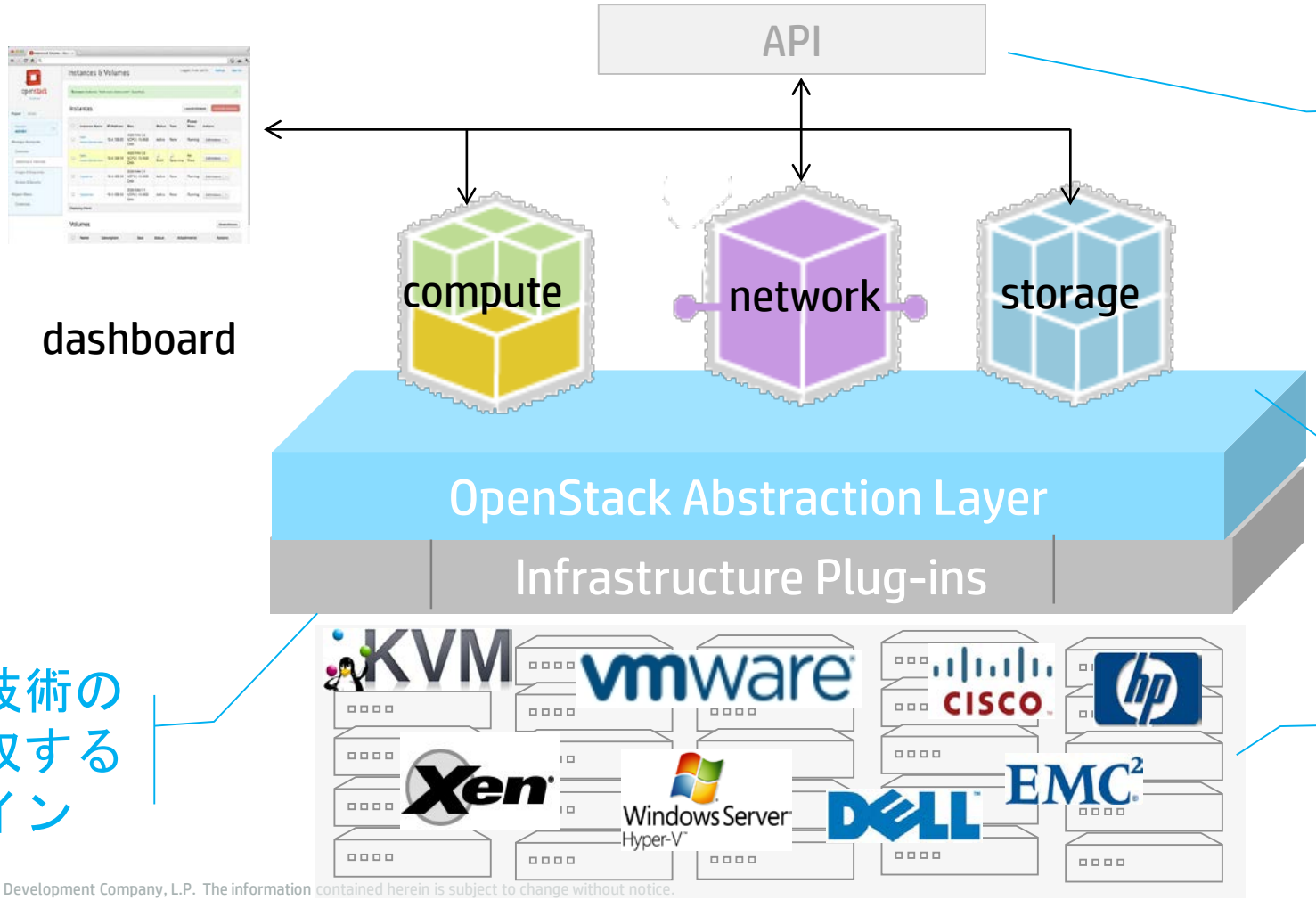
IT基盤の管理者のメリット

- 利用者ごとの個別対応が不要
- 運用の効率化と管理の向上
- ヘルプデスクの負荷軽減
- 統合によるコスト削減効果

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

OpenStack概要

多様なインフラをサービス化する
特定ベンダの技術・仕様に縛られない＝”オープン・スタンダード”



標準化された
API

インフラを抽象化

ベンダと技術の
違いを吸収する
プラグイン

異種混在インフラ

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

OpenStack開発の経緯とロードマップ

Nebula(IaaS基盤)

2009年
独自のクラウドプラットフォームを開発・運営

Cloud Files(ファイルホスティング)

2008年
独自のクラウドファイルホスティングサービスを開発・運営

2010/7

ロードマップ

Austin	Bexer	Cactus	Diablo	Essex	Folsom	Grizzly	Havana	Icehouse	Juno	Kilo
2010/10	2011/2	2011/4	2011/10	2012/4	2012/9	2013/4	2013/10	2014/4	2014/10	2015/04

Now !!

Next !!

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

OpenStack参加メンバー企業

Platinum Members

AT&T, Canonical, HP, IBM, Nebula, Rackspace, Red Hat, Inc., SUSE

Gold Members

Aptira, CCAT, Cisco, Cloudscaling, Dell, DreamHost, eNovance, Ericsson, Hitachi, Huawei, Intel, Juniper Networks, Mirantis, Morphlabs, NEC, NetApp, Piston Cloud, VMware, Yahoo!

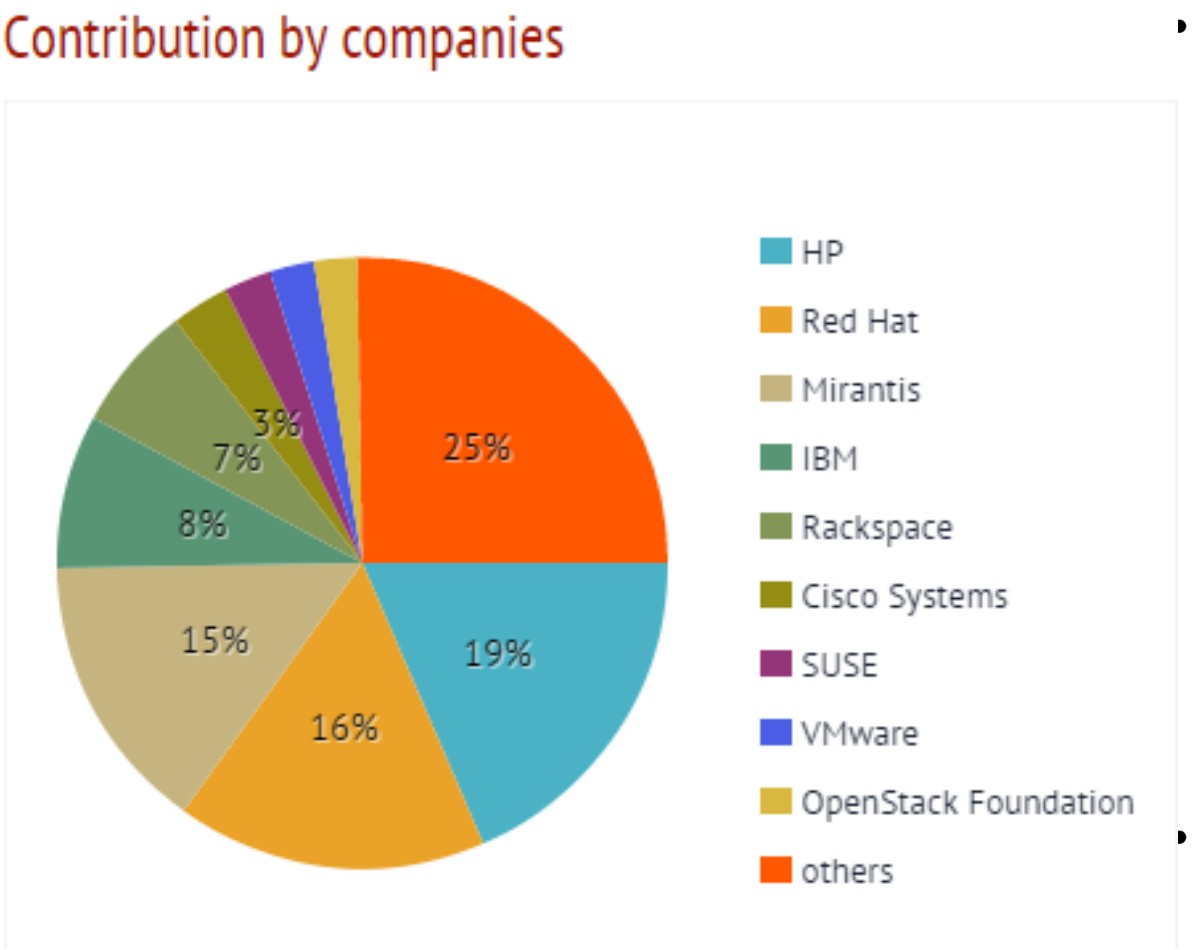
http://www.openstack.org/foundation/companies/

Corporate Sponsors

Adaptive Computing, Alcatel Lucent, Anuta Networks, Arista Networks, Avaya, Blue Box, Brocade, Bull, Canadian Web Hosting, Cloudwatt, Comcast, Coraid, Cumulus Networks, Cyan Inc., EMC, F5 Networks, Fujitsu, Fusion-io, Gale Technologies, Gridcentric, hastexo, Hortonworks, Internap, IO, iWeb, KIO Networks, liquidweb, Metacloud, Nexenta, Nexus, Nokia Solutions and Networks, Numergy, Nutanix, NxtGen, Oracle, Parallels, PayPal, RightScale, RiverMeadow Software, Scality, Seagate, Servergy, Smartscale Systems, StackLabs, SwiftStack Inc, Thales Group, Transcend Computing, UnitedStack Inc., VCE, Western Digital Corporation, Wind River, Xemeti, ZTE Corporation

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

HPのOpenStackへの貢献と経験



http://stackalytics.com/ (2015/01/19)

- HPメンバーが開発をリード(PTL)するプロジェクト
  - Horizon (Dashboard)
  - Neutron (Network)
  - Trove (DBaaS)
  - Ironic (Bare-Metal)
  - TripleO (Provisioning)
  - Designate (DNSaaS)
  - Infrastructure (CI Platform)
  - Quality Assurance (CI)
  - Olso (共通ライブラリ)
- 2011/9から3年以上 OpenStackベースのパブリッククラウドを運用してきた経験を活かす、還元する

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

オープンソース版OpenStack 機能一覧 (Juno)

正式機能

Phase	プロジェクト	機能
Integrated	Nova	サーバ管理
	Swift	オブジェクトストレージ
	Glance	システムイメージ管理
	Cinder	ブロックストレージ管理
	Keystone	統合認証
	Horizon	ダッシュボード
	Neutron	ネットワーク管理
	Ceilometer	メータリング/モニタリング
	Heat	オーケストレーション
	Trove	DB as a Service
	Sahara	Hadoopクラスタ管理

次のリリースで正式機能を目指す機能

Phase	プロジェクト	機能
Incubation	Ironic	ベアメタルサーバ管理
	Marconi	メッセージキュー
	Designate	DNS管理

新たに提案された機能（成熟を待ってIncubationへ）

Phase	プロジェクト	機能
External	Graffiti	メタデータ管理
	Barbican	キー管理
	Murano	サービスカタログ管理
	Staccato	データ転送

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

「オンプレミス型インフラ」と「クラウド型インフラ」

オンプレミス型(「ペット」型)

- ペットのように1台1台のサーバに名前をつける
- 24時間365日手厚く面倒を見る
- 調子が悪くなったらすぐ手当てをする

クラウド型(「家畜」型)

- 家畜のように名前ではなく、番号で扱う
- どのサーバも区別は付けない/付かない
- 調子が悪くなったら、別のサーバを起動

"CERN Data Center Evolution", Gavin McCance, 2012

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Hadoop-as-a Serviceとは

ビッグデータの広がり

デジタルデータ、非構造化データが指数関数的に増大

毎秒の Tweet 数
97,000

オーディオ

YouTube 上のビデオ数
1億本以上

ソーシャル
メディア

ビデオ

毎月 Facebook で共有
されるコンテンツ数
300億

SharePoint ユーザー
1億人以上
> 30%の増加率

イメージ

Email

2010年携帯電話数
50億台

テキスト

ドキュメント

米国における
紙ドキュメント数**4兆**
増加率 22%

一日のメール数
2,940億通

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

ビッグデータ処理に対するニーズ

背景

- eコマース、広告業の競争激化
- クリック数、ページビュー急減
- センサーデータの活用

顧客と商品の相関関係、類似度を見たい
ユーザーの行動パターン予測

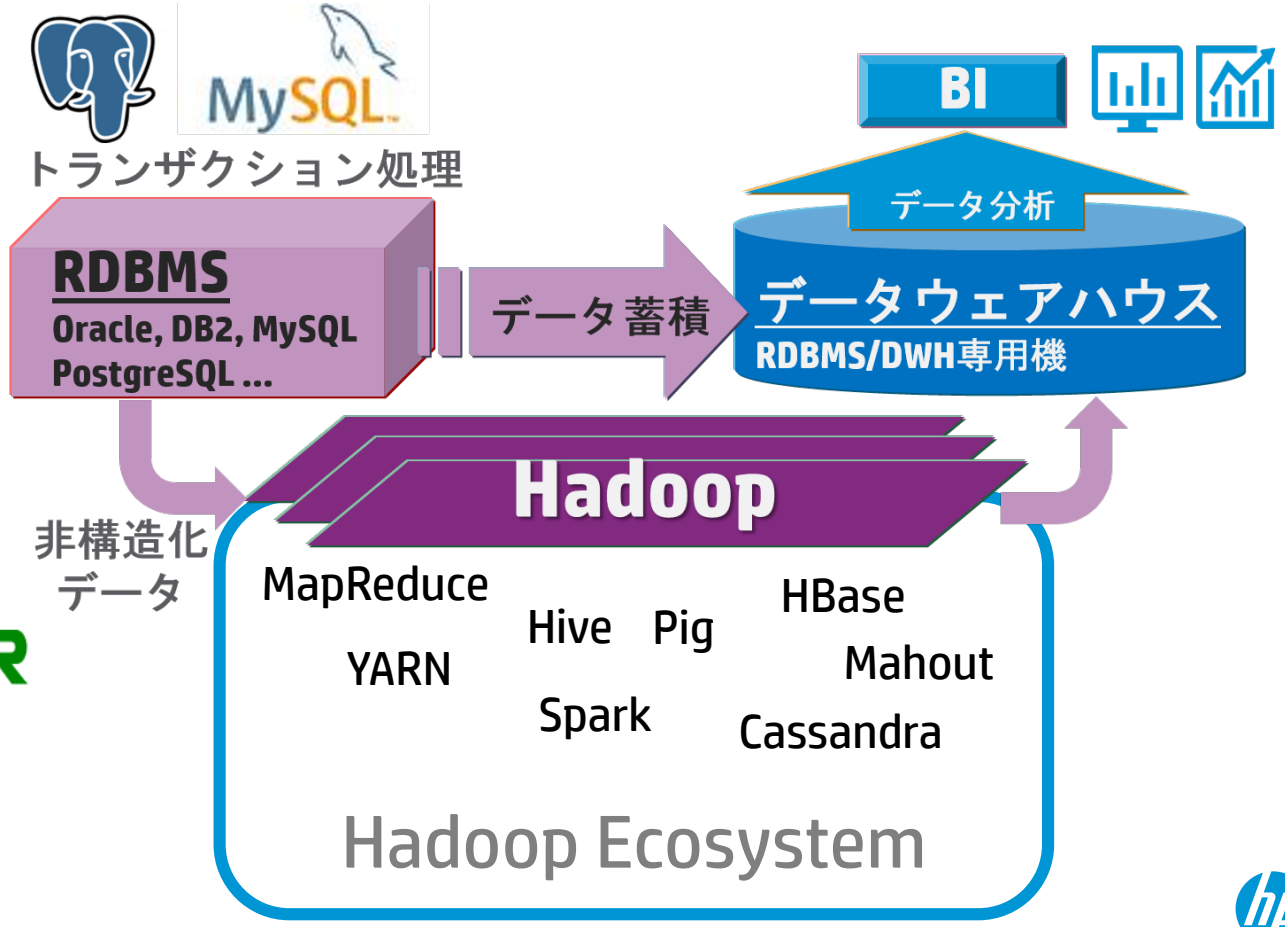
- レコメンド
- ユーザー行動分析
- ログ分析
- Web Page Indexing

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Hadoopとは
非構造化データに対する、大規模分散データ処理フレームワーク

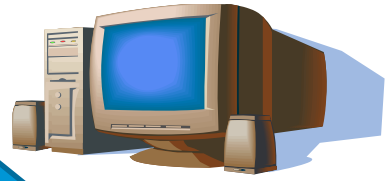
Google File SystemやMapReduce論文発表（2004年）

Doug Cutting氏：Hadoopを開発
Apache Hadoopプロジェクトとして公開



© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Hadoopシステム構成



クライアント/データソース

データの読み書き
処理の依頼

Hadoop クラスタ

Name Node / Job Tracker

役割：メタデータの保持、Data Nodeの状態管理、Job管理
構成：HAソフトウェアによる冗長化

Map/Reduce ：複数のサーバで分散処理することで高性能を実現

HDFS(Hadoop Distributed File System)
ファイルを分割して複数のサーバに複製して保持することで冗長性を担保

Data Node / Task Tracker

データの保持と分散処理を、スケールアウト構成で実現

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Hadoop-as-a-Serviceとは

オンプレミスのHadoopクラスタ基盤の課題への1つの回答

オンプレミスのHadoopクラスタ基盤の課題

Hadoopクラスタ基盤の構築/資産化

- Hadoop専用のクラスタシステムが必要
- 有用な処理結果が得られるかは実行しなければならない

Hadoopクラスタ基盤の運用

- クラスタの運用管理の負荷が継続的に発生
- クラスタのスケールアウト計画

サービスとしてのHadoopクラスタ

クラウドリソースを利用して構築

- 必要な時にクラウド上のリソースからHadoopクラスタを素早く構築
- データ処理結果を簡単に得られる

使用時にのみHadoopクラスタ基盤を運用

- クラスタの運用管理は、ほぼ無くなる
- 必要に応じてクラウド上のリソースを追加

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Sahara概要

Sahara概要

OpenStack上でHadoopを動かす！

Saharaは、OpenStackで構成したクラウド上で、ユーザがHadoopクラスタを簡単にデプロイし、また管理可能にすることを目的として開発されたコンポーネントです。

主なユースケース：

- 開発やQAのための、OpenStack上へのHadoopクラスタの素早いデプロイ
- 汎用目的のOpenStack IaaSクラウドの未使用コンピュートパワーの活用
- アドホックな”Analytics as a Service”(分析サービス)や、突発的な分析ワークロード
  - Amazon Elastic MapReduce (EMR - http://aws.amazon.com/jp/elasticmapreduce/)と同様なユースケース

利用上の利点

- 気軽にデータ処理を試すことができる
- 必要な時に必要なサイズのHadoopクラスタを用意し、利用することができる
- 利用しない場合はHadoopクラスタを削除し、リソースを他の用途に利用できる
- REST APIを利用した自動Hadoopクラスタ構築・ジョブ実行が可能

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Sahara概要

Saharaでできること

① Hadoopクラスタ基盤構築

- Hadoopクラスタ管理
  - OpenStackクラウド上でのHadooクラスタ基盤の構築・削除・管理が可能
- Plug-inベース
  - 様々なディストリビューション（コミュニティ版、Hortonworks、Cloudera、MapR）を利用可能
- クラスタ・スケーリング
  - 構築したHadoopクラスタのサイズを簡単に増減できる

② ジョブ実行 - Elastic Data Processing

- ジョブ実行
  - Saharaで構築したHadoopクラスタに対し、ジョブを実行することが可能
- Swift、HDFS連携
  - 処理対象データをSwiftまたは外部のHDFSに配置し、利用することが可能
  - また、処理結果を出力することも可能
- ジョブ管理
  - OpenStack Horizonからジョブ管理を行うことが可能

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

① Hadoopクラスタ基盤構築

1. 利用するプラグインを選択

- 利用するディストリビューション、バージョンを選択

2. イメージレジストリの登録

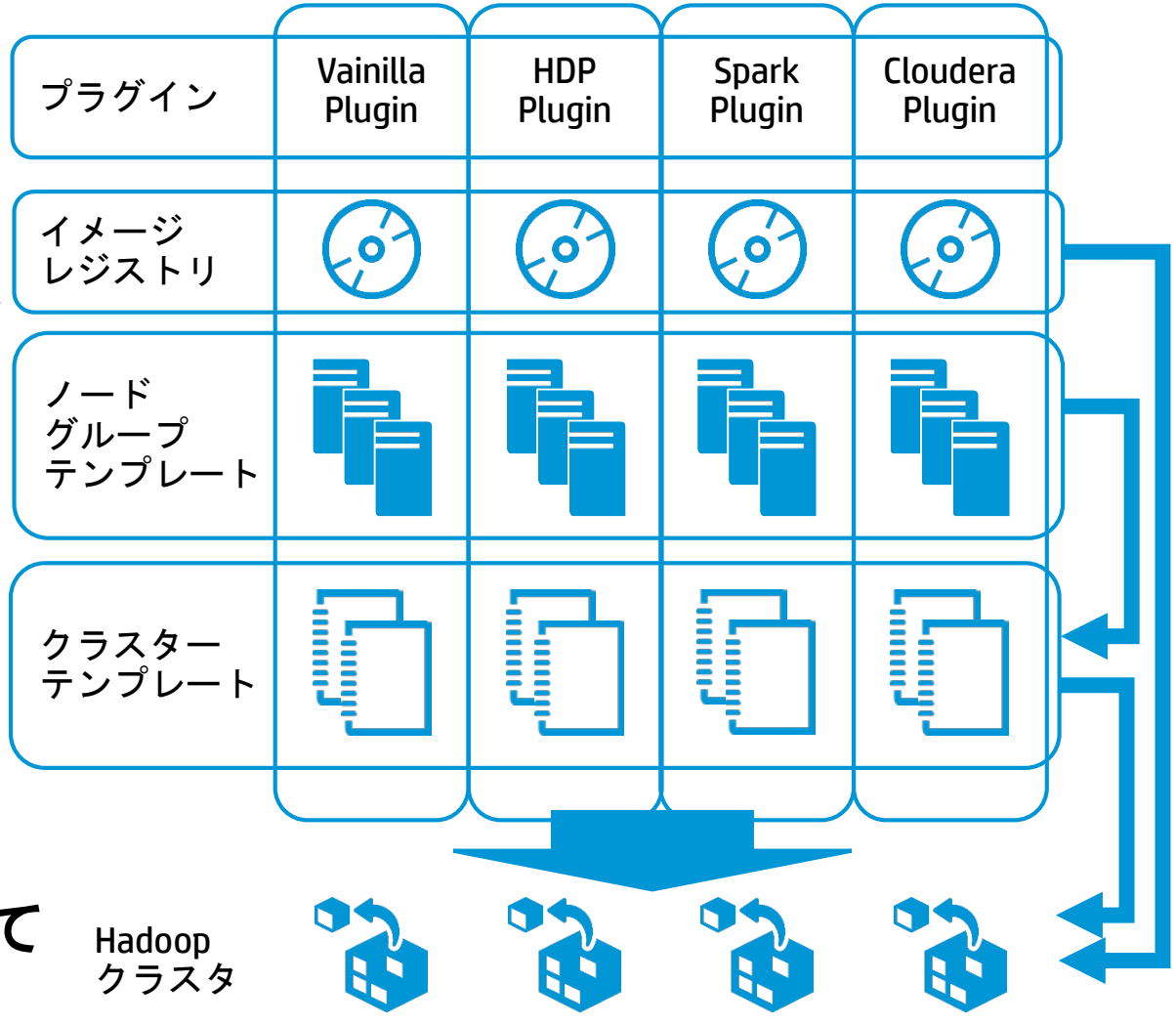
- Glanceに登録したSahara用イメージとプラグインを紐付けてSaharaに登録

3. ノードグループテンプレート作成

- Hadoopのマスタノード、スレーブノード等、ノードの役割毎のテンプレートをプラグインと紐付けて作成

4. クラスタテンプレート作成

- 1つ以上のノードグループテンプレートからクラスタのテンプレートを作成



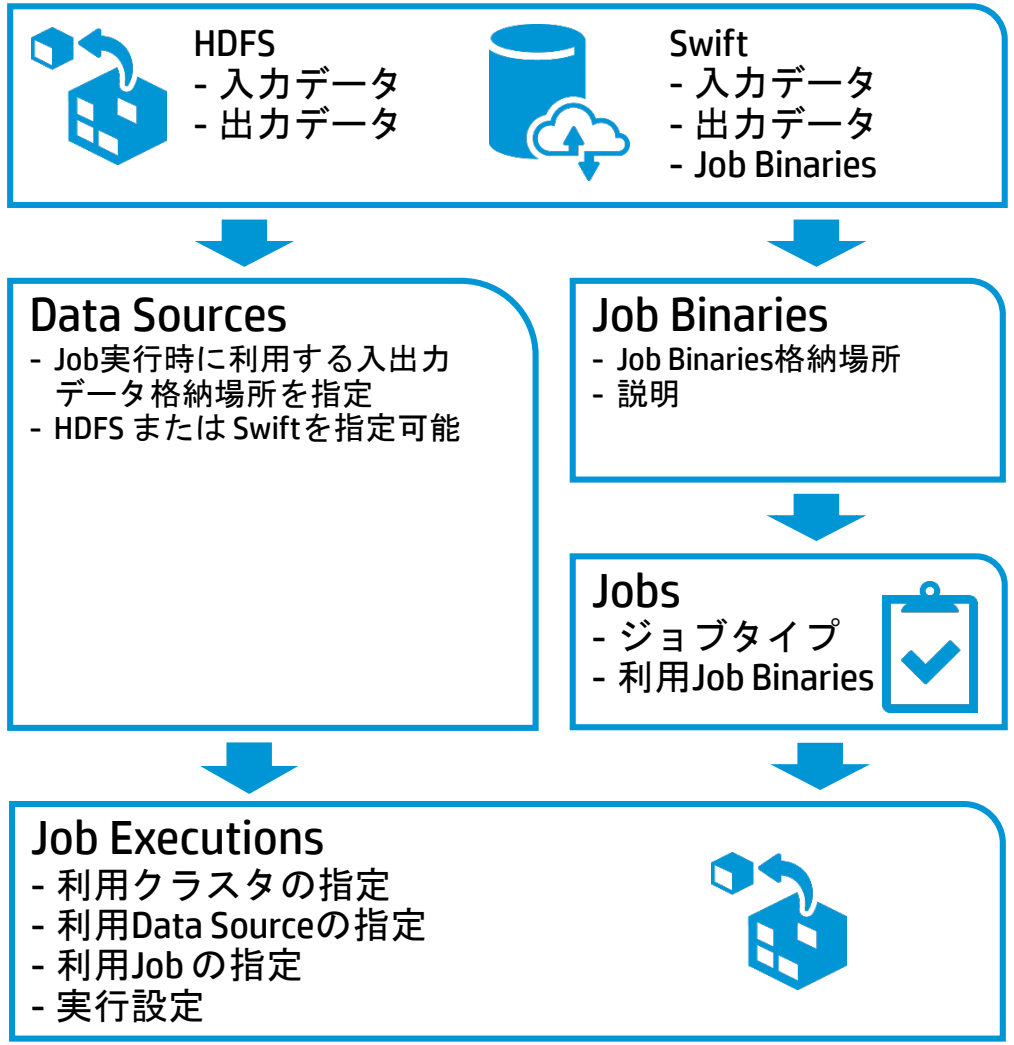
クラスターテンプレートを利用して Hadoopクラスタをデプロイ！

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

② ジョブ実行 – Elastic Data Processing

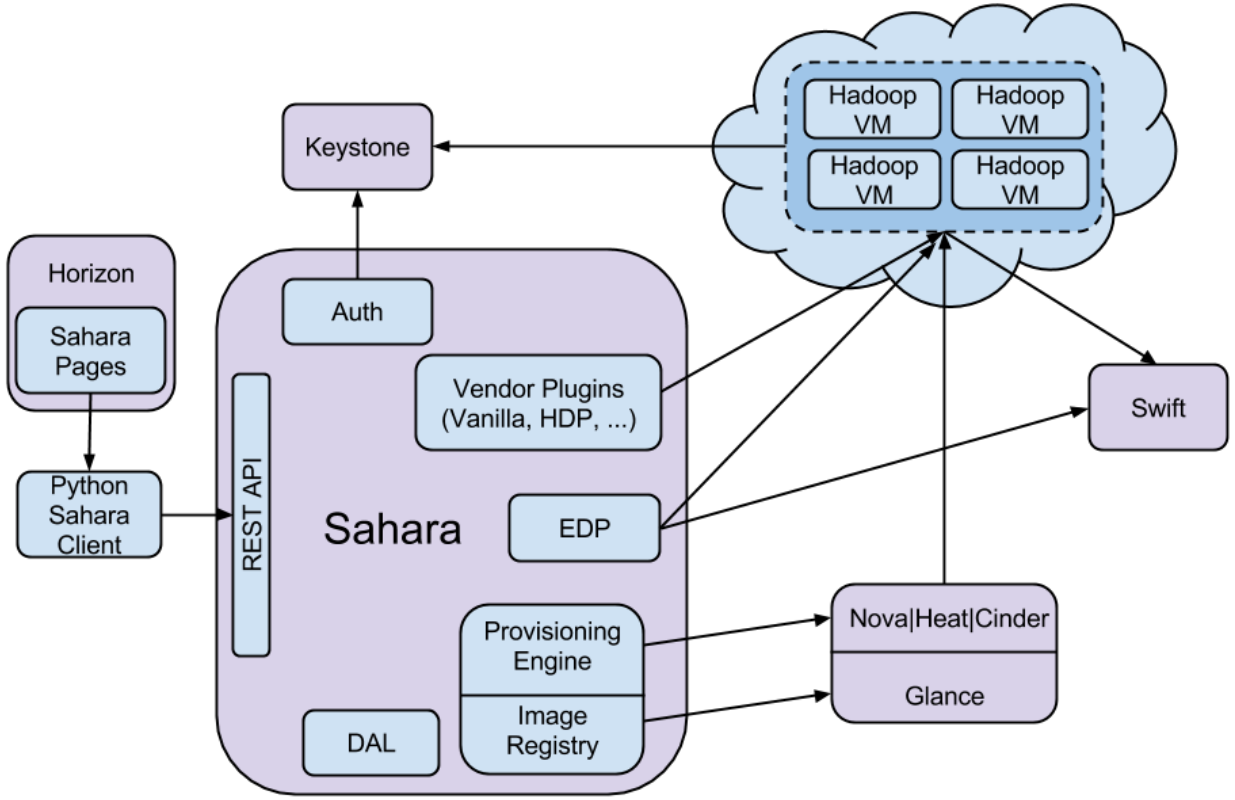
1. 入力データおよび処理プログラム配置
   - 外部のHDFSまたはSwift上に入力データを配置
   - 必要に応じ、処理プログラムも配置
2. データソース設定
   - 処理対象データへのアクセスURLをデータソースとして設定
   - 処理結果の格納場所も同様に設定
3. ジョブバイナリ登録
   - データ処理プログラム（スクリプトやJavaライブラリ等）を設定
4. ジョブ設定
   - ジョブタイプ（Pig, Hive, MapReduce,Spark等）とジョブバイナリを指定し、ジョブを設定します。

入出力用のデータソースおよびジョブを指定して、ジョブを実行！



© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Saharaのアーキテクチャ



コンポーネント	概要
Sahara Pages	HorizonでのSahara UIを提供
Sahara Client	CLI
REST API	Saharaの機能を提供するAPI
Auth	Keystoneと連携し、認証認可を行う
Venders Plugins	プロビジョニングしたVM上のHadoopの設定および起動に責任を持つ
EDP(Elastic Data Processing)	ジョブ実行のスケジューリングと管理に責任を持つ
Provisioning Engine / Image Registry	Nova, Heat, Cinder, Glanceと通信に責任を持つ
DAL(Data Access Layer)	内部モデルをDB永続化する

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Hadoopクラスタデモ

デモ

- Tips
  - DataSourceの指定
    - Swiftの場合
      - EDPを実行するローカルのOpenStackのSwift上のコンテナを利用することができる
      - 指定URL は以下のパターン
        - <コンテナ名>.sahara/<オブジェクト名>
        - 例） demo.sahara/input.tar
    - HDFSの場合
      - Httpfs を利用してアクセスする

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Appendix
Sahara参考情報

参考情報

- プロジェクトサイト
  - http://docs.openstack.org/developer/sahara/index.html
- Git
  - https://github.com/openstack/sahara
  - EDP実行サンプル
    - https://github.com/openstack/sahara/tree/master/etc/edp-examples
- Red Hat Enterprise Linux OpenStack Platfrom 5
  - **OpenStack のデプロイメント: 実習環境 (手動設定) 第11章 OPENSTACK SAHARA のインストール**
  - https://access.redhat.com/documentation/ja-JP/Red_Hat_Enterprise_Linux_OpenStack_Platform/5/html/Installation_and_Configuration_Guide/chap-OpenStack_Sahara_Installation.html
- OpenStack Data Processing (Sahara) Juno Release Features Demo
  - http://youtu.be/zHYfLNJ7ncI
  - 上記で利用していたサンプル
    - https://github.com/openstack/sahara/tree/master/etc/edp-examples

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Sahara用イメージ

- Vanilla Plugin用
  - http://docs.openstack.org/developer/sahara/userdoc/vanilla_plugin.html
  - Hadoop 1.2.1 および 2.4.1 が存在

- HDP Plugin用
  - http://docs.openstack.org/developer/sahara/userdoc/hdp_plugin.html
  - Hortonworks社のHadoopディストリビューション

- Spark Plugin用
  - http://docs.openstack.org/developer/sahara/userdoc/spark_plugin.html
  - CDH4 HDFS + Apache Sparck

- Cloudera Plugin用
  - http://docs.openstack.org/developer/sahara/userdoc/cdh_plugin.html
  - CDH 5.2ベース
  - Disk Image Builderでイメージを作成する必要がある

© Copyright 2014 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.

Thank you!