

クラウドストレージの基礎知識

～ ビッグデータに最適なクラウドストレージの特徴と S3 エコシステムのメリットとは？ ～

クラウドストレージの分野が急速に成長しています。代表的なクラウドストレージサービス、Amazon S3 は 2013 年には、2 兆個を超えるオブジェクト数に達したことを発表しました。日本においても、ビッグデータに対応できるクラウドストレージサービスが本格スタートしています。企業ではプライベート型のクラウドストレージ構築も始まっています。本ペーパーでは、クラウドストレージの特徴と S3 エコシステムのメリットについて解説します。

1 あらためてビッグデータについて考えてみましょう

ある日突然に多種多様な非構造の大量データがシステムを急襲する。「ビッグデータ」は、そんな現象のことです。インターネットをモバイル環境で利用し、誰もが時間と場所の制約無しにソーシャルメディア等から情報を発信します。人間だけではなく、駅の自動販売機までもが購入者の性別や年齢を見分け、時間の記録データを送信し続けます。このような機械と機械間の通信 (M2M) もビッグデータの膨張に拍車をかけています。

ビッグデータへの関心の高まりは、オープンソースの分散処理ソフトウェアである Hadoop が重要な役割を果たしたと言えるでしょう。Hadoop 登場以前には、大量データ処理は高価で高性能なマシンを使うか、何時間もの処理が普通でした。Hadoop は分散処理ソフトウェアを身近なものとなりました。

Hadoop のような分散処理ソフトウェアは、Google、Amazon、Facebook、Twitter といった代表的なインターネット企業で従来から利用されています。顧客の Web サービスにおける利用や行動、購買履歴の大量データを分析し、嗜好に合ったサービスや商品の推薦をビジネスにするという利用事例が目まぐるしく目立っています。

この「Hadoop は大量データ処理に適した身近な技術」であることと、「大量データ分析はビジネスになる」ことが結びついて語られ

ることが増えました。最近では「大量データ分析=ビッグデータ」と広く認識されているように見受けられます。しかし、いざデータ分析を始めてみると、実はそれほど手元にデータは無かったという現実もあるようです。それゆえに、ビッグデータはバズワードであり、一過性の流行語であるといった声も聞こえ始めています。

しかし、そもそも「ビッグデータ」とは、インターネット技術者達が、ビッグウェーブ (大波) のようにデータが押し寄せる現象を語ったことに始まります。多種多様で膨大なデータを迅速に「書き込み」、「読み出す」ことが求められる時代の到来を示唆しているのです。



2 ビッグデータに対応できるクラウドストレージの特徴

すでにビッグデータに直面しているのがクラウドサービスです。Amazon は、クラウドストレージサービスの Amazon S3 (Simple Storage Service) が保存するオブジェクト数が 2013 年には 2 兆個を超えたと発表しました。保存データ量、データの読み出し、リクエスト数に課金し、1GB あたり月額平均で 10 円を下回ると言われています。ストレージするデータ量に制限はなく、利用者は使った

分だけを支払う従量制サービスです。日本においては 2011 年 9 月、ニフティクラウドストレージが従量制で無制限にストレージできるクラウドサービスを本格スタートして以来、NTT コミュニケーションズの「Biz ホスティング Cloud n Object Storage」等、多くのクラウド事業者がクラウドストレージサービスを提供しています。この Amazon S3 やニフティクラウドストレージは、ビッグデータに対応

できるクラウドストレージサービスです。このクラウドストレージの特徴を、次の3つの点から説明しましょう。

- (1) オブジェクトストレージ
- (2) 分散処理ソフトウェア
- (3) ストレージの仮想化

(1) クラウドストレージはオブジェクト構造

ビッグデータに対応するクラウドストレージの1つの特徴は、オブジェクトストレージが主流であるという点です。このオブジェクトストレージのオブジェクト構造では、ひとつひとつのオブジェクトに `http://s3.cloudian.com/abc123def...` といった URL が ID として付与されます。このオブジェクト構造の利点を、一般的に利用されているファイルシステムのファイル構造と比較しながら説明します。

a) オブジェクト構造はデータ移動が容易

ファイル構造とは、図2に示すように階層構造です。階層構造をもつ ID の頂点にはサーバー名があり、この階層のいずれかにコンテンツが保存されています。階層構造は、人間の思考にマッチしています。地球があって、日本があって、東京があって、渋谷があって、私のオフィスがあるといった具合に、ごく自然な思考の流れで目標に到達できます。

しかし、このファイル構造はビッグデータを扱うとなると厄介な問題が生じてきます。たとえば、大量のコンテンツを保存した結果、全体容量がひっ迫したとします。それらのコンテンツを他のサーバーに移動すると ID も変更されます。小規模なグループ利用であれば、保存場所の変更は軽微な通知で済むでしょう。しかし、数千人、数万人規模が利用すると、その変更には大きなインパクトを生じます。このように ID に場所情報が含まれていると、データの移動が簡単ではありません。オブジェクト構造であれば、すべての ID の関係性はフラットです。保存場所とは無関係にコンテンツを移動できます。このデータ移動が容易であることがクラウド環境には求められるのです。

b) WAN を介した複数データセンター対応

クラウドでは、ひとつのデータセンターに障害があってもサービス停止やデータ紛失がないよう、複数のデータセンターで相互にバツ

クアップします。フラットな構造であれば、複数データセンターにまたがる複雑な階層構造を避けることができます。また、オブジェクト構造のデータ入出力に用いる HTTP はインターネットの標準プロトコルです。広域ネットワーク(WAN)を経由した複数データセンター間でのデータのやり取りも、安全におこなうことができます。

c) 大量のファイル数を扱える

ファイル構造では、大規模になると取り扱えるファイル数に制限が生じることや、性能に課題が生じることがあります。一方、オブジェクト構造であれば、同様の制約はありません。このことは、Amazon S3 が2兆個を超えるオブジェクトを格納していると報告していることから明らかです。

d) その他のメリット

その他、オブジェクト構造であれば、オブジェクト毎に属性情報などのメタ情報を追加できることや、オブジェクト毎に定めた条件での利用やアクセスの許可を付与することができるといったメリットもあります。

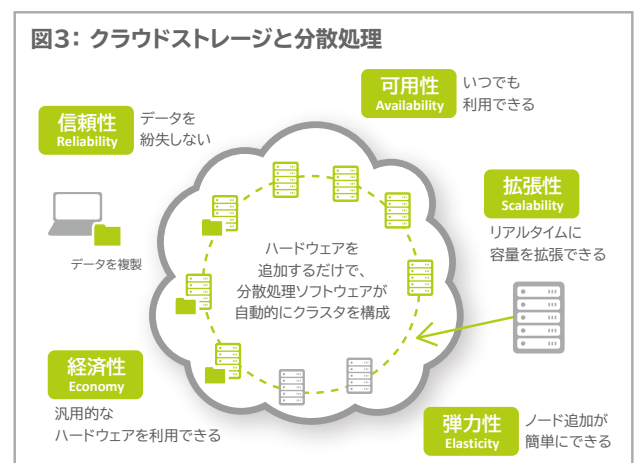
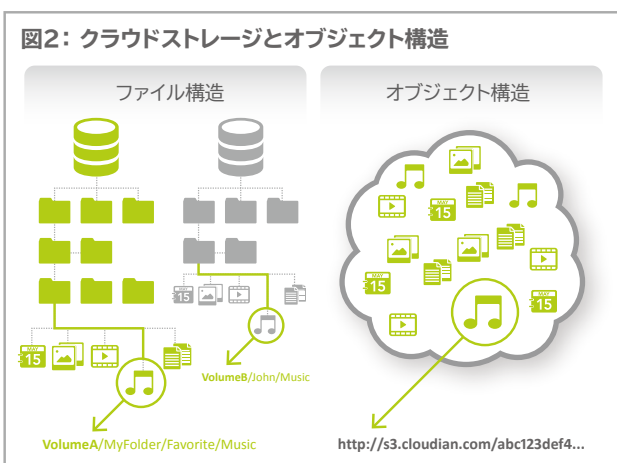
(2) 分散処理ソフトウェアによる拡張

クラウドストレージの2つの特徴としては、多数の汎用的なサーバーを同時並行して使う分散処理ソフトウェア技術(NOSQLデータベースや分散ファイルシステム)が利用されているという点です。

突発的(バースト的)にデータが大波となり押し寄せるのがビッグデータです。将来を見越した容量のストレージ装置を用意しても、1、2年で足りなくなってしまうかもしれません。装置に障害が起きることもあるでしょう。耐用年数に応じて置換しなければなりません。ハードウェアに依存したストレージ装置の場合、いずれかのタイミングでデータを移行する必要が生じます。それには人手と時間がかかります。

そのためクラウドストレージでは、汎用的なサーバーを利用し、次のような特性をもつソフトウェアで分散処理をおこないます。

- 一部のサーバーに障害があってもサービスは停止しない(可用性)
- サーバーを追加することで全体容量を拡張できる(拡張性)
- サーバー追加や縮退に対し自動的に対応し、負荷をリバランスする(弾力性)



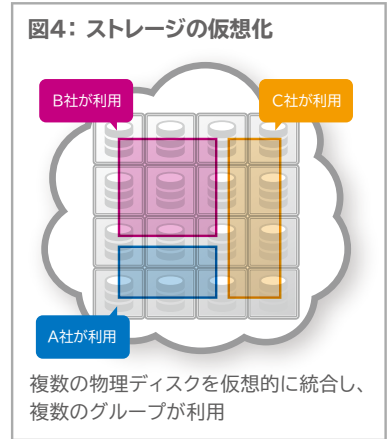
- 汎用的なサーバーを使い、経済的にシステムを構築できる（経済性）
- 複数サーバーにデータを複製し、システムとしてデータ紛失を防ぐ（信頼性）

(3) ストレージの仮想化によりマルチテナントで共有

クラウドストレージの3つの特徴はストレージを仮想化することです。パブリック型のクラウドストレージは、企業内利用に限定されるプライベート型とは異なり、独立した多数のグループがシステムを共有します。この独立したグループに、個別の設備を用意して物理的なストレージ資源を割当てるクラウドストレージサービスも存在はしています。これらの多くは、定額で利用量制限のあるサービスとして提供されています。この場合、クラウドサービスの提供側は、利用者毎に装置を購入する設備投資が必要です。利用者側は、結果として利用しない設備容量のコストも負担することになります。つまり、クラウドが得意とする規模のメリットを双方が享受することが難しくなります。

「使った分だけを支払う」クラウドサービスのためには、ストレージの仮想化により、多数の利用者がひとつのストレージ資源を共有できることが不可欠です。

ストレージの仮想化とは、多数の物理ディスクをあたかもひとつのストレージ資源であるかのように仮想的に統合すると同時に、複数グループ（マルチテナント）で共有できるようにすることです。このストレージの仮想化により、クラウドサービスの提供側は、動的に空いている設備容量を利用者に割当てることができ、利用者側にとっては使った容量分だけのコストを負担すれば良くなります。



クラウドストレージはAPIを提供する

さて、ここで少し、今後のクラウドストレージの方向性を考えてみます。パブリック型のクラウドストレージの利用に際して、政府が閲覧権限を持つ国にデータセンターがあると、セキュリティ面が不安との声を聞くことがあります。こういった声は日本に限らず欧州やアジア各国でも聞こえ始めています。その一方で、ビッグデータは従来のITシステムの想定外であり、その対応は多くの企業にとって解決すべき大きな課題です。そのため、パブリック型に留まらず、企業が監視できるデータセンターでデータを自社管理できるプライベート型、又はパブリック型と組み合わせたハイブリッド型といったクラウドストレージが広く求められるようになります。

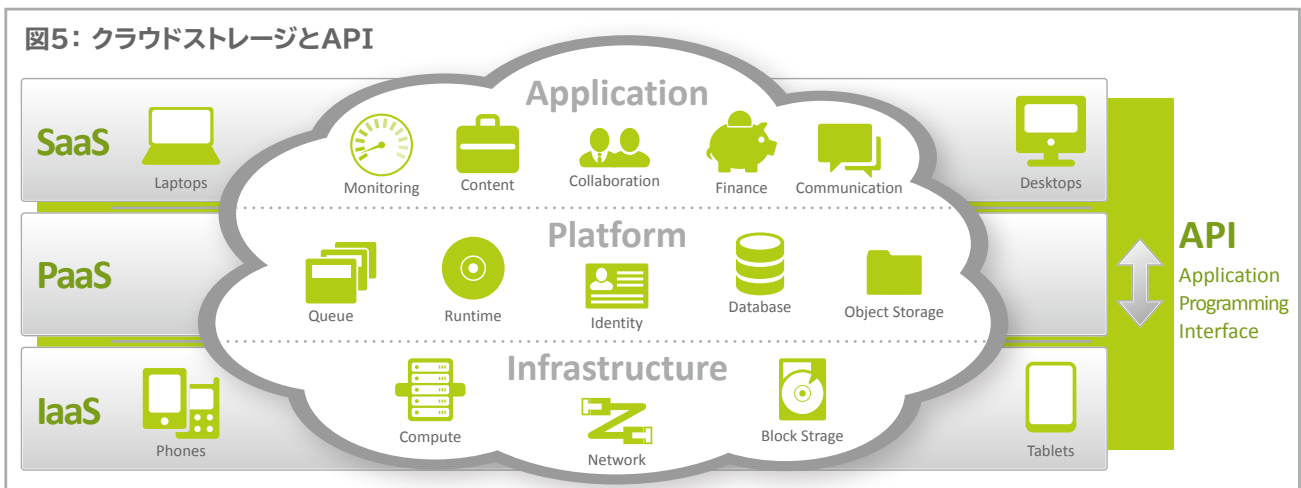
クラウドストレージを構築、導入すると考えた場合、「クラウドストレージはAPIを提供する」という重要な特徴に注目することが必要です。

図5は、クラウド・コンピューティング俯瞰図です。一般的にクラウドサービスは、アプリケーションを提供するサービスをSaaS (Software as a Service)、プラットフォームを提供するサービスを

PaaS (Platform as a Service)、インフラを提供するサービスをIaaS (Infrastructure as a Service) と分類されます。

この分類に従うと、クラウドストレージはアプリケーションにインフラを提供するIaaSかプラットフォームを提供するPaaSです。このクラウドストレージとアプリケーション間をつなぐインターフェースがAPI (Application Program Interface) であり、APIを提供することがクラウドストレージの特徴と言えます。

ひとつの例として、最近注目を集めているDropbox (ドロップボックス) という「オンラインストレージサービス (ファイル共有サービスとも呼ばれます)」があります。このDropboxは大量の写真や数MBのファイルも、オブジェクトやフォルダーのIDである <https://www.dropbox.com/s/xxx...> といったリンクのアドレスを送受するだけで仲間と共有できます。Amazon S3はこのDropboxにクラウドストレージサービスを提供しており、このDropboxとAmazon S3をつなぐインターフェースがAPIです。



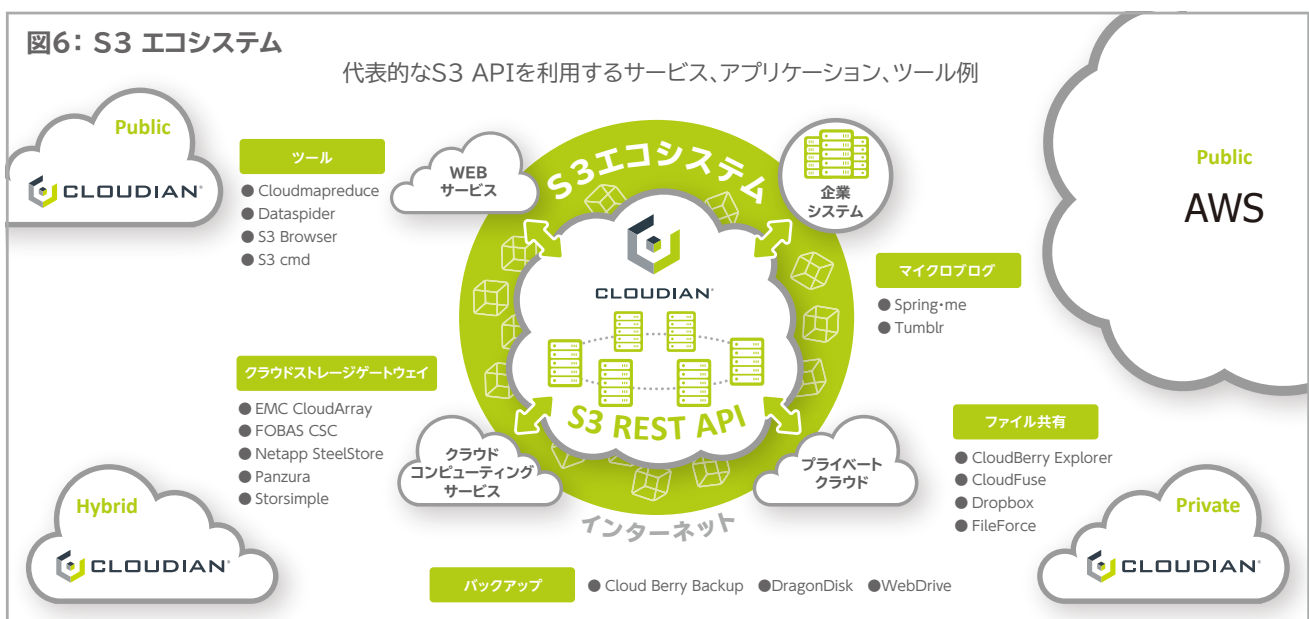
4 S3エコシステムのメリットを活用

このクラウドストレージのAPIは、パブリック型であれ、プライベート型であれ、クラウドストレージを利用するアプリケーションとの接点です。より多くの利用者にとって利便性の高いことが重要です。それは、APIに関するドキュメントや情報が整備され、開発者が多数存在し、そのAPIを利用するアプリケーション、ツール、サービス、アプライアンスなどが豊富に揃っていることです。また、事実上の標準となっているクラウドストレージのAPIであれば、1つのアプリケーションでパブリック型とプライベート型の相互運用も容易になります。クラウドストレージ毎に独自アプリケーションなどを開発する時間やコストも不要になります。つまり、クラウドストレージの生態系、すなわちエコシステムが重要なのです。

クラウドストレージにおいて、そのエコシステムがもっとも充実し

ているのは、Amazon S3を中心としたS3エコシステムであると言えるでしょう。S3 APIの仕様は広く公開され、S3を利用できるサービスやアプリケーションは数百種類が公表されています。実際のところ、千種類を超えるとも言われています。

S3に準拠するAPIを提供するクラウドストレージであれば、利用者は、同じアプリケーションによるAWS (Amazon Web Service) との相互利用も可能です。同時に豊かなS3エコシステムのメリットを存分に享受できます。近い将来、クラウド間を相互接続して、相互にバックアップしあうインタークラウドの動きも想定されるでしょう。そのようなインタークラウドの動きに際しても、S3 APIに準拠したクラウドストレージ間の相互接続であれば、差分を調整する中間層も不要であることから、スムーズに進むことが期待できます。



まとめ

ビッグデータは、多種多様で膨大なデータを迅速に「書き込み」、「保存し」、「読み出す」という課題への対処が求められる時代の到来を示唆しています。このビッグデータに直面するクラウドストレージの特徴は、(1) クラウド環境に適したオブジェクト構造、(2) 高い拡張性、弾力性、信頼性を備える分散処理、(3) ストレージの仮想化

による統合と複数グループによる共有であると言えます。今後、パブリック型に加えプライベート型の構築も活発化します。その際、豊かなエコシステムを持つS3 APIに準拠したクラウドストレージを構築、導入することで多くのメリットを享受できます。

クラウドファンについて

日本と米国を開発拠点とするクラウドファンは、パブリッククラウド、プライベートクラウド、オンプレミス環境でハイブリッドに活用できるSDS (Software Defined Storage: ソフトウェア定義ストレージ) である「Cloudian HyperStore」ソフトウェア製品及びアプライアンス製品により提供しています。国内外大手プロバイダー、エンタープライズが採用するCloudian HyperStoreは、複数データセンター間を含み、データ複製・分散配置によるデータ保護をしながら、汎用サーバ2台からペタバイト超級にまで経済的に、柔軟にスケールアウトします。統計・課金・管理機能も実装済みであり短期間に利用開始できます。



クラウドファン株式会社

www.cloudian.jp | info@cloudian.com