社会経済システム学会大会　京都産業大学
分科会Ⅰ　「イノベーション・エコシステム」
2013年10月10月26日（土）10：00～11：40

報告要旨：大規模データ解析におけるミドルウェア開発・利用の普及プロセス
―　開発コミュニティと市場構造からみる米国のHadoopのケース・スタディ　―

早稲田大学デジタル・ソサエティ研究所　田中絵麻

はじめに

ITシステムにおけるハードウェア部分がコモディティ化するなか、OSとアプリケーションを統合するミドルウェア部分のソフトウェアの重要性が増している。一方で、企業はミドルウェアをクローズドにすることで競争力の向上をはかる誘因があるため、ミドルウェア部分でのベンダーロックインが発生しうる。しかし、米国においては、クラウド・サービス[1]やビッグデータ解析サービスにおけるミドルウェアのオープンソース化が行われており、結果的に特定の企業の技術に依存するロックインの回避と市場の活性化が起きている。米国でどのように、自己強化的（特定企業）なベンダーロックインではなく、コミュニティ強化的（ソフトウェア産業発展）なオープンソース・ソフトウェア開発が行われ、イノベーションの普及が展開するのか、そのプロセスと要因について、2007年にオープンソース化されたHadoopの拡大プロセスのケース・スタディから考察する。

1　Hadoopの開発体制

Hadoopの元となったオープンソース・ソフトウェア（以下、OSS）は、2002年に開始された検索エンジンNutchである。その後、2003年と2004年にGoogleが分散ファイルシステムと大規模データ処理技術に関する論文を公表、Nuchがこれらの技術を取り込む形で2006年から本格的にHadoopの開発が行われた。その開発には、Yahoo!の社員もかかわっており、Yahoo!の検索エンジンの改良がその目的であった。その後、OSSのウェブサーバ・ソフトウェアのApacheを開発しているApache Software Foundationのプロジェクトの一つとして、2007年9月にバージョン0.14.1が公開された。2013年9月現在、Hadoopプロジェクトは、Apacheの749あるプロジェクトのなかでも大規模なプロジェクトに成長した。表1は、情報が提供されている509のプロジェクトの開発者数（Committer）を集計したものである。Hadoopは、18名と報告されているが、実際のプロジェクトページでは、66名がリストアップされている。

[1] クラウド・サービスにおけるオープンソース・ソフトウェアのミドルウェアの事例としては、NASAのNebulaを引き継いだOpenStackが挙げられる。

表1　Apacheにおける開発人数毎のプロジェクト数（2013年9月現在）

開発者数	1	2	3	4	5	6	7	8	9	10	11～15	16～25	26～35
プロジェクト数	345	61	25	12	18	4	5	4	5	8	11	8	3

出所：Apacheより集計.

表2は、Hadoopの開発者の所属別の人数を集計したものである。Hortonworksとは、2011年にYahoo!のHadoop開発チームが独立して設立されたベンチャー企業である。HadoopとYahoo!の開発者を合計すると全体の48%を占めており、実質的にYahoo!が主導するOSS開発プロジェクトだと言えよう。

表2　Hadoopの開発者の所属別人数（開発者数に占める割合）

Hortonworks	Yahoo!	Cloudera	Facebook	IBM	LinkedIn	その他※
22（33%）	10（15%）	8（12%）	5（8%）	3（5%）	3（5%）	15（23%）

※2名がHuawei、InMobi、Microsoft、WANdisco。1名がGetopt、INRIA、UC Berkeley、無所属。

出所：Apacheより集計.

2　Hadoopエコシステムの拡大

Hadoopの特徴は、Googleの検索システムの構成要素をクローン化したミドルウェアである点だと考える。開発の中核を担ってきたYahoo!は、検索サービスの市場シェアで2002年にGoogleに逆転されている。そのため、Hadoopを支援する競争上のインセンティブがある。なお、Googleでは、検索サービスの中核技術を自社利用するというクローズド戦略を基盤として、無料サービスと広告のマッチングにより収益モデルを確立した。ただし、一方で、Googleは、学術分野へのサポートを重視しており、上述の2本の論文以外にも多数の論文をオンラインで公開している。

Hadoopのベータ版の公開後、米国を起点とした「ビックデータ・ブーム」が発生した。学術情報の検索サービスであるGoogle Scholarで「big data」を検索すると、2010年代に急速に大規模データを指す用語として拡大した事がわかる。

表　3　"big data"のヒット件数の推移

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
229	294	319	383	416	558	735	1,020	1,860	7,020	11,000

出所：Google Scholar, 2013年9月現在.

同様に、Google Trendで"big data"の検索件数のボリューム推移を見ると、ピーク時のヒット件数を100とした場合、2010年7月までの検索ボリュームは1であったものがそ

の後急速に拡大、2012年6月には40、2013年9月のボリュームが100となっている。

このように、「big data」というワードが、大規模データ解析サービスを指すようになったのはごく近年の事であり、その一つの要因として、米国におけるO'Reilly Mediaによる関連イベントが影響していると考える。なお、O'Reilly Mediaでは、2004年から「Web 2.0」に関するイベントを開催、このワードも人口に膾炙した。O'Reilly Mediaは、2011年初頭から、「データは新たな石油である」としてビッグデータに関するイベント・シリーズ「Strata Conference」を開催している。第2回目のイベントでは、IT開発者やCIO、データサイエンティストといったIT分野のスペシャリストやジャーナリスト等の参加者が2,000名を超えた。表4は、Strata Conferenceのスポンサー、展示企業数の集計結果である。ロンドンでの開催時の企業数は少ないものの、サンタクララ、ニューヨークでの開催は回を重ねる毎に増加している。なお、2014年の数は2013年9月現在である。

表4　Strata Conferenceのスポンサー、展示企業数

2011S	2012S	2012N	2012L	2013S	2013N	2013L	2014S
28	42	75	13	80	95	15	27

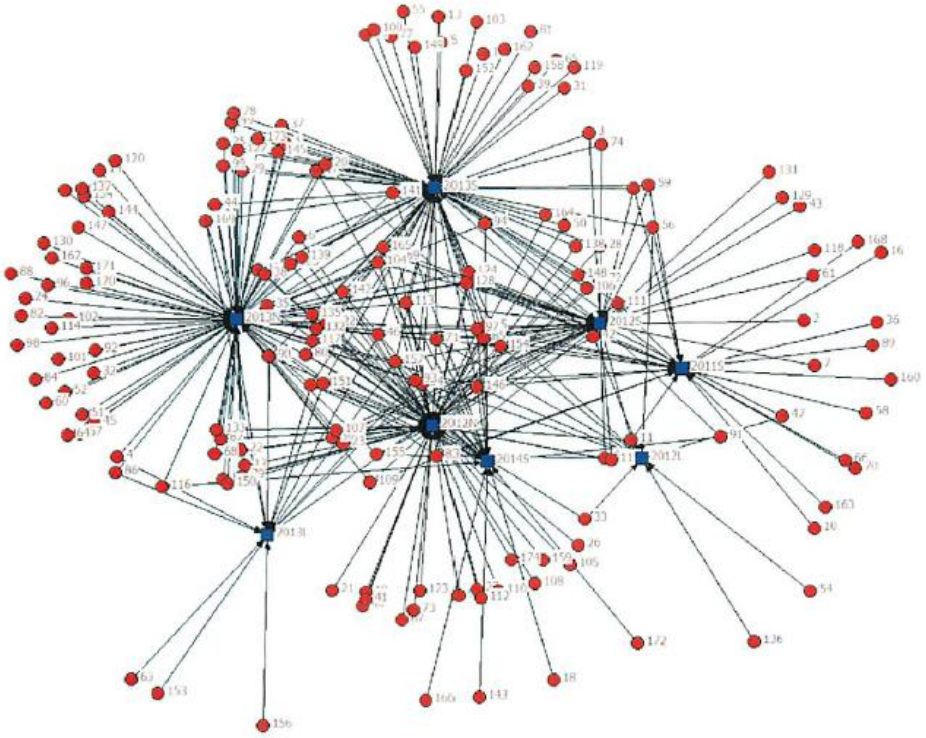
Sはサンタクララ開催、Nはニューヨーク開催、Lはロンドン開催。

出所：Strata Conferenceサイト.

図1は、参加企業がどのイベントに参加したかを示す二部グラフである。二部グラフでは、イベントと参加者の関係とイベントへの中核的な参加者を視覚化することができる。調査した8イベントでの総参加企業数は174社で、8イベントに参加していたのは、Splunk（下図146）のみで、7イベントがMapR（93）、6イベントが、Cloudera（34）、Hortonworks（71）、IBM（75）、Impetus（76）、Karmasphere（85）、Microsoft（97）、Tableau Software（154）、Teradata Aster（157）であった。HadoopのOSS開発の中核企業のHortonworks、Clouderaのほか、商用Hadoopベンダーや Hadoop関連製品を取り扱うSplunk、Karmasphere、MapR、Impetus、Tableau Software、Teradata Asterが名を連ねている。これらの企業は二部グラフの中心に集まっている。一方、Googleは、いずれのイベントでも、スポンサーもしくは展示参加企業とはなっていない。

なお、Hadoop関連の企業は、大規模分散データベース関連のソリューションや製品についてそれぞれ相互に組み込み可能にしている場合が多い。そのため、Hadoop関連製品、企業全体を指して「Hadoopエコシステム」と呼称している。なお、Google も、HadoopをGoogle Cloud Platformで利用可能としている。

図1　Strata Conferenceへのスポンサー、展示参加企業の二部グラフ



おわりに

本稿では、Hadoopのケース・スタディから、OSSを中核とした関連製品・サービス拡大の背景には、①Yahoo!とGoogleの競争関係、②Googleの事業戦略と学術分野におけるポリシー、③Hadoopを巡る協調的企業間関係があることを指摘した。また、この過程において、米国初のビッグデータ・ブームが起きており、大規模分散データ解析サービス、ソリューションへの認知も向上したと言える。また、Yahoo!という、経営資源の一部（開発者）を共有し、公共財化（OSSとしてのHadoop）する中核的企業（イネーブラー）が、ロックイン回避というインセンティブ（市場競争下）持っていたことで、企業間関係によるエコシステムの形成が推進されたといえる。なお、本稿の分析は一事例にとどまっており、一般化には理論的検討、さらなるケース分析や実証分析が必要である。

参考文献：Fransman, Martin (2010) *The New ICT Ecosystem: Implications for Policy and Regulation*, Cambridge University Press.

マルコ・イアンシティ（2007）『キーストーン戦略 イノベーションを持続させるビジネス・エコシステム』翔泳社.　（なお、本要旨で参照した参考記事等の出所は紙面の関係から発表時に表示する）