Me bifurque no GitHub

A tabela do ecossistema do Hadoop

Esta página é um resumo para manter o controle de projetos relacionados ao Hadoop, com foco no ambiente FLOSS.

	Sistema de Arquivos Distribuídos	
Apache HDFS	O Hadoop Distributed File System (HDFS) oferece uma maneira de armazenar arquivos grandes em várias máquinas. O Hadoop e o HDFS foram derivados do papel do sistema de arquivos do Google (GFS). Antes do Hadoop 2.0.0, o NameNode era um ponto único de falha (SPOF) em um cluster HDFS. Com o Zookeeper, o recurso HDFS High Availability resolve esse problema fornecendo a opção de executar dois NameNodes redundantes no mesmo cluster em uma configuração ativa / passiva com um hot standby.	1. hadoop.apache.org 2. Google FileSystem - GFS Paper 3. Cloudera Por que o HDFS 4. Hortonworks Por que o HDFS
GlusterFS da Red Hat	O GlusterFS é um sistema de arquivos de armazenamento anexado à rede escalável. O GlusterFS foi desenvolvido originalmente pela Gluster, Inc., e depois pela Red Hat, Inc., após a compra da Gluster em 2011. Em junho de 2012, o Red Hat Storage Server foi anunciado como uma integração suportada comercialmente do GlusterFS com o Red Hat Enterprise Linux. Gluster File System, conhecido agora como Red Hat Storage Server.	1. www.gluster.org 2. Plugin do Red Hat Hadoop
Quantcast File System QFS	O QFS é um pacote de software de sistema de arquivos distribuídos de código-fonte aberto para MapReduce de grande escala ou outras cargas de trabalho de processamento em lote. Ele foi projetado como uma alternativa ao HDFS do Apache Hadoop, destinado a oferecer melhor desempenho e custo-benefício para clusters de processamento em larga escala. Ele está escrito em C ++ e possui gerenciamento de memória de tamanho fixo. O QFS usa correção de erros Reed-Solomon como método para garantir acesso confiável aos dados. A codificação Reed – Solomon é amplamente usada em sistemas de armazenamento em massa para corrigir os erros de rajada associados a defeitos de mídia. Em vez de armazenar três versões completas de cada arquivo, como o HDFS, resultando na necessidade de três vezes mais armazenamento, o QFS precisa de	1. Site do QFS 2. GitHub QFS 3. HADOOP-8885

	apenas 1.5x a capacidade bruta, pois divide os dados em nove unidades de disco diferentes.	
Sistema de arquivos Ceph	Ceph é uma plataforma de armazenamento de software livre projetada para apresentar armazenamento de objetos, blocos e arquivos a partir de um único cluster de computador distribuído. Os principais objetivos do Ceph são ser completamente distribuídos sem um único ponto de falha, escalável ao nível do exabyte e livremente disponível. Os dados são replicados, tornando-os tolerantes a falhas.	1. Site do sistema de arquivos Ceph 2. Ceph e Hadoop 3. HADOOP-6253
Sistema de arquivos Lustre	O sistema de arquivos Lustre é um sistema de arquivos distribuído de alto desempenho destinado a ambientes maiores de rede e alta disponibilidade. Tradicionalmente, o Lustre é configurado para gerenciar dispositivos de disco de armazenamento de dados remotos em uma Rede de Área de Armazenamento (SAN), que é dois ou mais dispositivos de disco conectados remotamente que se comunicam por meio de um protocolo SCSI (Small Computer System Interface). Isso inclui Fibre Channel, FCoE (Fibre Channel over Ethernet), SAS (Serial Attached SCSI) e até iSCSI. Com o Hadoop HDFS, o software precisa de um cluster dedicado de computadores nos quais executar. Mas as pessoas que executam clusters de computação de alto desempenho para outras finalidades geralmente não executam o HDFS, o que os deixa com um monte de poder de computação, tarefas que certamente poderiam se beneficiar de um pouco de redução de mapa e nenhuma maneira de colocar esse poder em funcionamento Hadoop. A Intel notou isso e, na versão 2.5 de sua distribuição Hadoop que lançou discretamente na semana passada, adicionou suporte ao Lustre: o Software de Distribuição Intel® HPC para Apache Hadoop *, um novo produto que combina o software Intel Distribution para Apache Hadoop com Intel® Enterprise Edition para o software Lustre. Esta é a única distribuição do Apache Hadoop que é integrada ao Lustre, o sistema de arquivos paralelo usado por muitos do mundo.	1. wiki.lustre.org/ 2. Hadoop com Lustre 3. Intel HPC Hadoop
Alluxio	O Alluxio, o primeiro sistema virtual de armazenamento distribuído centrado na memória do mundo, unifica o acesso a dados e une as estruturas de computação e os sistemas de armazenamento subjacentes. Os aplicativos só precisam se conectar ao Alluxio para acessar os dados armazenados em qualquer sistema de armazenamento subjacente. Além disso, a arquitetura centrada em memória do Alluxio	1. Site do Alluxio

	permite ordens de acesso de dados de magnitude mais rápida do que as soluções existentes. No ecossistema de big data, o Alluxio está entre estruturas de computação ou trabalhos, como Apache Spark, Apache MapReduce ou Apache Flink, e vários tipos de sistemas de armazenamento, como Amazon S3, OpenStack Swift, GlusterFS, HDFS, Ceph ou OSS. O Alluxio traz melhorias significativas de desempenho para a pilha; Por exemplo, o Baidu usa o Alluxio para melhorar o desempenho de análise de dados em 30 vezes. Além do desempenho, o Alluxio preenche novas cargas de trabalho com dados armazenados em sistemas de armazenamento tradicionais. Os usuários podem executar o Alluxio usando seu modo de cluster independente, por exemplo, no Amazon EC2, ou iniciar o Alluxio com o Apache Mesos ou o Apache Yarn. O Alluxio é compatível com o Hadoop. Isso significa que os programas Spark e MapReduce existentes podem ser executados no Alluxio sem nenhuma alteração de código. O projeto é de código aberto (Apache License 2.0) e é implantado em várias empresas. É um dos projetos de código aberto que mais cresce. Com menos de três anos de histórico de código aberto, o Alluxio atraiu mais de 160 colaboradores de mais de 50 instituições, incluindo Alibaba, Alluxio, Baidu, IBM, Intel, NJU, Red Hat, UC Berkeley e Yahoo. O projeto é a camada de armazenamento do Berkeley Data Analytics Stack (BDAS) e também parte da distribuição do Fedora.	
GridGain	GridGain é um projeto de código aberto licenciado sob o Apache 2.0. Uma das principais peças desta plataforma é o Acelerador In-Memory Apache Hadoop, que visa acelerar o HDFS e Mapear / Reduzir, trazendo dados e cálculos para a memória. Este trabalho é feito com o sistema de arquivos em memória compatível com GGFS - Hadoop. Para trabalhos intensivos de E / S, o GridGain GGFS oferece desempenho quase 100x mais rápido do que o HDFS padrão. Parafraseando Dmitriy Setrakyan da GridGain Systems falando sobre a GGFS em relação à Tachyon: • O GGFS permite leitura e gravação de / para o HDFS subjacente ou qualquer outro sistema de arquivos compatível com Hadoop com alteração de código zero. Essencialmente, o GGFS remove completamente a etapa de ETL da integração. • O GGFS tem a capacidade de escolher quais pastas permanecem na memória, quais pastas	1. GridGain site

	permanecem no disco e quais pastas são sincronizadas com o FS subjacente de maneira síncrona ou assíncrona. • O GridGain está trabalhando na adição do componente MapReduce nativo que fornecerá integração nativa completa do Hadoop sem alterações na API, como o Spark atualmente força você a fazer. Essencialmente, o GridGain MR + GGFS permitirá trazer o Hadoop total ou parcialmente na memória na forma Plug-n-Play sem quaisquer alterações na API.	
XtreemFS	O XtreemFS é um sistema de armazenamento de uso geral e cobre a maioria das necessidades de armazenamento em uma única implantação. É de código aberto, não requer hardware especial nem módulos de kernel, e pode ser montado em Linux, Windows e OS X. O XtreemFS é executado distribuído e oferece resiliência por meio de replicação. O XtreemFS Volumes pode ser acessado através de um componente FUSE, que oferece interação normal de arquivos com a semântica POSIX. Além disso, uma implementação da interface do Hadoops FileSystem está incluída, o que torna o XtreemFS disponível para uso com o Hadoop, Flink e Spark fora da caixa. O XtreemFS está licenciado sob a licença New BSD. O projeto XtreemFS é desenvolvido pelo Zuse Institute Berlin. O desenvolvimento do projecto é financiado pela Comissão Europeia desde 2006 no âmbito dos acordos de subvenção nº FP6-033576, FP7-ICT-257438 e FP7-318521,	1. Site do XtreemFS 2. Flink no XtreemFS . Spark XtreemFS
	Programação distribuída	
Apache Ignite	Apache Ignite In-Memory O Data Fabric é uma plataforma distribuída na memória para computação e transações em conjuntos de dados de grande escala em tempo real. Ele inclui um armazenamento na memória de valor-chave distribuído, recursos SQL, redução de mapas e outras computações, estruturas de dados distribuídas, consultas contínuas, subsistemas de mensagens e eventos, integração do Hadoop e Spark. O Ignite é construído em Java e fornece APIs .NET e C ++.	1. Apache Ignite 2. Documentação do Apache Ignite
Apache MapReduce	MapReduce é um modelo de programação para processar grandes conjuntos de dados com um algoritmo distribuído paralelo em um cluster. O Apache MapReduce foi derivado do documento Google MapReduce: Processamento Simplificado de Dados em Grandes Clusters. A versão atual do Apache MapReduce é construída sobre o Apache YARN	1. Apache MapReduce 2. Google MapReduce paper 3. Escrevendo aplicativos YARN

	Framework. YARN significa “Yet-Another-Resource-Negotiator”. É uma nova estrutura que facilita a criação de estruturas e aplicativos de processamento distribuídos arbitrários. O modelo de execução do YARN é mais genérico do que a implementação anterior do MapReduce. O YARN pode executar aplicativos que não seguem o modelo MapReduce, diferente do Apache Hadoop MapReduce original (também chamado MR1). O Hadoop YARN é uma tentativa de levar o Apache Hadoop além do MapReduce para processamento de dados.	
Porco Apache	O Pig fornece um mecanismo para executar fluxos de dados em paralelo no Hadoop. Inclui uma linguagem, Pig Latin, para expressar esses fluxos de dados. O Pig Latin inclui operadores para muitas das operações de dados tradicionais (junção, classificação, filtro, etc.), bem como a capacidade dos usuários de desenvolver suas próprias funções para leitura, processamento e gravação de dados. Pig é executado no Hadoop. Ele usa o sistema de arquivos Hadoop Distributed File, HDFS e o sistema de processamento do Hadoop, o MapReduce. O Pig usa o MapReduce para executar todo o processamento de dados. Ele compila os scripts Pig Latin que os usuários escrevem em uma série de um ou mais trabalhos MapReduce que são executados. Pig Latin parece diferente de muitas das linguagens de programação que você viu. Não há declarações if ou loops em Pig Latin. Isso ocorre porque as linguagens de programação tradicionais orientadas a objetos e procedimentos descrevem o fluxo de controle, e o fluxo de dados é um efeito colateral do programa. Pig Latin concentra-se no fluxo de dados.	1. pig.apache.org/ 2.Pig examples de Alan Gates
JAQL	O JAQL é uma linguagem de programação funcional e declarativa projetada especialmente para trabalhar com grandes volumes de dados estruturados, semi-estruturados e não estruturados. Como o próprio nome indica, um uso principal do JAQL é manipular dados armazenados como documentos JSON, mas o JAQL pode trabalhar em vários tipos de dados. Por exemplo, ele pode suportar XML, dados de valores separados por vírgula (CSV) e arquivos simples. Um recurso "SQL dentro do JAQL" permite que os programadores trabalhem com dados SQL estruturados enquanto empregam um modelo de dados JSON que é menos restritivo que seus equivalentes na Structured Query Language. Especificamente, o Jaql permite selecionar, juntar,	1. JAQL no Google Code 2. O que é o Jaql? pela IBM

	agrupar e filtrar dados armazenados no HDFS, da mesma forma que uma mistura de Pig e Hive. A linguagem de consulta do Jaql foi inspirada em muitas linguagens de programação e consulta, incluindo Lisp, SQL, XQuery e Pig. O JAQL foi criado por trabalhadores no IBM Research Labs em 2008 e liberado para código aberto. Embora continue sendo hospedado como um projeto no Google Code, onde uma versão para download está disponível sob uma licença do Apache 2.0, a principal atividade de desenvolvimento em torno do JAQL permaneceu centrada na IBM. A empresa oferece a linguagem de consulta como parte do pacote de ferramentas associado ao InfoSphere BigInsights, sua plataforma Hadoop. Trabalhando em conjunto com um orquestrador de fluxo de trabalho, o JAQL é usado no BigInsights para trocar dados entre tarefas de armazenamento, processamento e análise. Ele também fornece links para dados e serviços externos, incluindo bancos de dados relacionais e dados de aprendizado de máquina.	
Apache Spark	Estrutura de computação de cluster de análise de dados originalmente desenvolvida no AMPLab na UC Berkeley. O Spark se encaixa na comunidade de código aberto do Hadoop, com base no Hadoop Distributed File System (HDFS). No entanto, o Spark fornece uma alternativa mais fácil de usar ao Hadoop MapReduce e oferece desempenho até 10 vezes mais rápido do que os sistemas de geração anterior, como o Hadoop MapReduce, para determinados aplicativos. O Spark é uma estrutura para escrever programas distribuídos rapidamente. O Spark resolve problemas semelhantes, como o Hadoop MapReduce, mas com uma abordagem rápida na memória e uma API de estilo funcional limpa. Com sua capacidade de integração com o Hadoop e ferramentas integradas para análise de consultas interativas (Shark), processamento e análise de gráficos em larga escala (Bagel) e análise em tempo real (Spark Streaming), ele pode ser utilizado de forma interativa para processar e consultar rapidamente grandes conjuntos de dados. Para tornar a programação mais rápida, o Spark fornece APIs limpas e concisas no Scala, Java e Python. Você também pode usar o Spark interativamente a partir dos shells do Scala e do Python para consultar rapidamente grandes conjuntos de dados. O Spark também é o mecanismo por trás do Shark, um sistema de data warehousing totalmente compatível com o Apache	1. Apache Spark 2. Espelho de faísca no Github 3. RDDs - Papel 4. Spark: Cluster Computing ... - Paper Spark Research

	Hive que pode ser executado 100 vezes mais rápido que o Hive.	
Tempestade de Apache	Storm é um processador de eventos complexos (CEP) e uma estrutura de computação distribuída escrita predominantemente na linguagem de programação Clojure. É um sistema de computação distribuído em tempo real para processar fluxos de dados grandes e rápidos. Storm é uma arquitetura baseada no paradigma mestre-trabalhador. Portanto, um cluster Storm consiste principalmente de um nó mestre e trabalhador, com coordenação feita pelo Zookeeper. O Storm faz uso do zeromq (0mq, zeromq), uma biblioteca de rede avançada e incorporável. Ele fornece uma fila de mensagens, mas, ao contrário do middleware orientado a mensagens (MOM), um sistema 0MQ pode ser executado sem um intermediário de mensagens dedicado. A biblioteca foi projetada para ter uma API de estilo de soquete familiar. Originalmente criado por Nathan Marz e equipe do BackType, o projeto foi aberto depois de ser adquirido pelo Twitter. O Storm foi inicialmente desenvolvido e implantado no BackType em 2011. Após 7 meses de desenvolvimento, o BackType foi adquirido pelo Twitter em julho de 2011. O Storm foi aberto em setembro de 2011. A Hortonworks está desenvolvendo uma versão Storm-on-YARN e planeja concluir o nível básico integração em 2013 Q4. Esse é o plano da Hortonworks. O Yahoo / Hortonworks também planeja mover o código Storm-on-YARN de github.com/yahoo/storm-yarn para ser um subprojeto do projeto Apache Storm em um futuro próximo. O Twitter lançou recentemente um Hadoop-Storm Hybrid chamado "Summingbird". O Summingbird funde as duas estruturas em uma, permitindo aos desenvolvedores usar o Storm para processamento de curto prazo e o Hadoop para mergulhos profundos de dados. um sistema que visa mitigar as compensações entre o processamento em lote e o processamento de fluxo, combinando-os em um sistema híbrido.	1. Projeto Storm / 2. Storm-on-YARN
Apache Flink	O Apache Flink (anteriormente chamado de Stratosphere) apresenta abstrações poderosas de programação em Java e Scala, um tempo de execução de alto desempenho e otimização automática de programas. Ele tem suporte nativo para iterações, iterações incrementais e programas que consistem em grandes DAGs de operações. O Flink é um sistema de processamento de dados e	1. Página da incubadora Apache Flink 2. Site da estratosfera

	uma alternativa ao componente MapReduce do Hadoop. Ele vem com seu próprio tempo de execução, em vez de construir em cima do MapReduce. Como tal, pode funcionar de forma completamente independente do ecossistema do Hadoop. No entanto, o Flink também pode acessar o sistema de arquivos distribuídos (HDFS) do Hadoop para ler e gravar dados, e o gerenciador de recursos de próxima geração do Hadoop (YARN) para provisionar recursos de cluster. Como a maioria dos usuários do Flink está usando o Hadoop HDFS para armazenar seus dados, ele já envia as bibliotecas necessárias para acessar o HDFS.	
Apache Apex	O Apache Apex é uma plataforma de big data-in-motion baseada no Apache YARN de nível empresarial que unifica o processamento de fluxo, bem como o processamento em lote. Ele processa big data in-motion em um altamente escalonável, altamente performante, tolerante a falhas, com estado, seguro, distribuído e de fácil operação. Ele fornece uma API simples que permite aos usuários escrever ou reutilizar código Java genérico, diminuindo assim o conhecimento necessário para escrever aplicativos de big data. A plataforma Apache Apex é complementada pelo Apache Apex-Malhar, que é uma biblioteca de operadores que implementa funções comuns de lógica de negócios necessárias para clientes que desejam desenvolver aplicativos rapidamente. Esses operadores fornecem acesso ao HDFS, S3, NFS, FTP e outros sistemas de arquivos; Kafka, ActiveMQ, RabbitMQ, JMS e outros sistemas de mensagens; MySql, Cassandra, MongoDB, Redis, HBase, CouchDB e outros bancos de dados junto com conectores JDBC. A biblioteca também inclui uma série de outros padrões de lógica de negócios comuns que ajudam os usuários a reduzir significativamente o tempo necessário para entrar em produção. A facilidade de integração com todas as outras tecnologias de big data é uma das principais missões do Apache Apex-Malhar. O Apex, disponível no GitHub, é a principal tecnologia na qual a oferta comercial da DataTorrent, DataTorrent RTS 3, juntamente com outras tecnologias, como a ferramenta de processamento de dados chamada dtIngest, são baseadas.	1. Apache Apex de DataTorrent 2. Apache Apex página principal 3. Apache Apex Proposal
Pigpen Netflix	PigPen é map-reduce para o Clojure que compila para o Apache Pig. Clojure é o dialeto da linguagem de programação Lisp criada por Rich Hickey, portanto, é	1. PigPen no GitHub

	uma linguagem funcional de propósito geral e é executada nos mecanismos Java Virtual Machine, Common Language Runtime e JavaScript. No PigPen não há funções especiais definidas pelo usuário (UDFs). Defina as funções do Clojure, anonimamente ou nomeadas, e use-as como você faria em qualquer programa do Clojure. Essa ferramenta é aberta pela Netflix, Inc., o provedor americano de mídia de streaming de Internet sob demanda.	
AMPLab SIMR	O Apache Spark foi desenvolvido pensando no Apache YARN. No entanto, até agora, tem sido relativamente difícil executar o Apache Spark em clusters do Hadoop MapReduce v1, ou seja, clusters que não possuem o YARN instalado. Normalmente, os usuários teriam que obter permissão para instalar o Spark / Scala em algum subconjunto das máquinas, um processo que poderia ser demorado. O SIMR permite que qualquer pessoa com acesso a um cluster do Hadoop MapReduce v1 execute o Spark fora da caixa. Um usuário pode executar o Spark diretamente sobre o Hadoop MapReduce v1 sem nenhum direito administrativo e sem ter o Spark ou o Scala instalado em nenhum dos nós.	1. SIMR no GitHub
Facebook Corona	"A próxima versão do Map-Reduce" do Facebook, baseada no próprio fork do Hadoop. A atual implementação do Hadoop da técnica MapReduce usa um único rastreador de trabalho, o que causa problemas de dimensionamento para conjuntos de dados muito grandes. Os desenvolvedores do Apache Hadoop estão criando seu MapReduce de próxima geração, denominado YARN, que os engenheiros do Facebook observaram, mas desconsiderou devido à natureza altamente personalizada da implantação do Hadoop e HDFS pela empresa.A Corona, como o YARN, gera vários rastreadores de trabalho (um para cada trabalho, no site da Corona caso).	1. Corona no Github
Apache REEF	O Apache REEF ™ (Framework de Execução do Avaliador Retentor) é uma biblioteca para o desenvolvimento de aplicativos portáteis para gerenciadores de recursos de cluster, como Apache Hadoop ™ YARN ou Apache Mesos ™. O Apache REEF simplifica drasticamente o desenvolvimento desses gerenciadores de recursos por meio dos seguintes recursos: • Fluxo de Controle Centralizado: O Apache REEF transforma o caos de um aplicativo distribuído em eventos em uma única máquina, o	1. Site Apache REEF

Job Driver. Os eventos incluem alocação de contêiner, inicialização de tarefas, conclusão e falha. Para falhas, o Apache REEF faz todo o esforço para tornar a 'Exceção' real lançada pela Tarefa disponível para o Driver.
- Tempo de execução da tarefa: o Apache REEF fornece um tempo de execução da tarefa chamado avaliador. Os avaliadores são instanciados em cada contêiner de um aplicativo REEF. Os avaliadores podem manter os dados na memória entre Tarefas, o que permite pipelines eficientes no REEF.
- Suporte para vários gerenciadores de recursos: os aplicativos Apache REEF são portáveis para qualquer gerenciador de recursos suportado com o mínimo de esforço. Além disso, novos gerenciadores de recursos são fáceis de suportar no REEF.
- API .NET e Java: Apache REEF é a única API para gravar aplicativos YARN ou Mesos no .NET. Além disso, um único aplicativo REEF é livre para misturar e combinar tarefas escritas para .NET ou Java.
- Plugins: O Apache REEF permite que plugins (chamados de "Serviços") aumentem seu conjunto de recursos sem adicionar um inchaço ao núcleo. O REEF inclui muitos Serviços, como comunicações baseadas em nome entre comunicações do grupo inspiradas em tarefas MPI (Broadcast, Reduce, Gather, ...) e entrada de dados.

Apache Twill	O Twill é uma abstração sobre o Apache Hadoop® YARN que reduz a complexidade do desenvolvimento de aplicativos distribuídos, permitindo que os desenvolvedores se concentrem mais em sua lógica de negócios. A Twill usa um modelo simples baseado em encadeamentos que os programadores Java vão achar familiar. O YARN pode ser visto como uma malha computacional de um cluster, o que significa que aplicativos YARN como o Twill serão executados em qualquer cluster do Hadoop 2. YARN é um aplicativo de código aberto que permite que o cluster do Hadoop se transforme em uma coleção de máquinas virtuais. O Weave, desenvolvido pela Continuuity e inicialmente instalado no Github, é um aplicativo de software livre complementar que usa um modelo de programação semelhante aos encadeamentos Java, facilitando a gravação de aplicativos distribuídos. Para remover um conflito com um projeto de nome semelhante no Apache, chamado	1. Incubadora Apache Twill

	"Weaver", o nome de Weave foi alterado para Twill quando foi movido para a incubação do Apache. O Twill funciona como um proxy escalonado. O Twill é uma camada de middleware entre o YARN e qualquer aplicativo no YARN. Quando você desenvolve um aplicativo Twill, o Twill lida com APIs no YARN que se assemelham a um aplicativo multiencadeado familiar ao Java. É muito fácil criar aplicativos distribuídos multiprocessados em Twill.	
Damballa Parkour	Biblioteca para desenvolver programas MapReduce usando o LISP como linguagem Clojure. O Parkour visa fornecer integração profunda do Clojure para o Hadoop. Programas usando o Parkour são programas Clojure normais, usando funções padrão do Clojure em vez de novas abstrações de framework. Programas usando o Parkour também são programas completos do Hadoop, com acesso completo a absolutamente tudo possível no Java Hadoop MapReduce.	1. Projeto Parkour GitHub
Apache Hama	Projeto de código aberto de alto nível Apache, permitindo que você faça análises avançadas além do MapReduce. Muitas técnicas de análise de dados, como algoritmos de aprendizado de máquina e gráficos, requerem cálculos iterativos, onde o modelo Bulk Synchronous Parallel pode ser mais eficaz que o MapReduce "simples".	1. site da Hama
Datasalt Pangool	Um novo paradigma MapReduce. Uma nova API para tarefas de RM, em nível superior ao Java.	1.Pangool 2.GitHub Pangool
Apache Tez	Tez é uma proposta para desenvolver um aplicativo genérico que pode ser usado para processar DAGs de tarefas complexas de processamento de dados e é executado nativamente no Apache Hadoop YARN. Tez generaliza o paradigma MapReduce para uma estrutura mais poderosa baseada na expressão de cálculos como um grafo de fluxo de dados. O Tez não se destina diretamente aos usuários finais - na verdade, ele permite que os desenvolvedores criem aplicativos para o usuário final com desempenho e flexibilidade muito melhores. Tradicionalmente, o Hadoop tem sido uma plataforma de processamento em lote para grandes quantidades de dados. No entanto, há muitos casos de uso para o desempenho quase em tempo real do processamento de consultas. Há também várias cargas de trabalho, como Machine Learning, que não se encaixam no paradigma MapReduce. Tez ajuda o Hadoop a abordar esses casos de uso.	1. Apache Tez Incubator 2. página Hortonworks Apache Tez
Apache DataFu	O DataFu fornece uma coleção de tarefas do Hadoop MapReduce e funções em linguagens de alto nível	1. DataFu Apache Incubator

	baseadas nele para executar a análise de dados. Ele fornece funções para tarefas estatísticas comuns (por exemplo, quantiles, amostragem), PageRank, sessionization de fluxo e operações de set e bag. O DataFu também fornece tarefas do Hadoop para processamento incremental de dados no MapReduce. O DataFu é uma coleção de Pig UDFs (incluindo PageRank, sessionization, operações de conjunto, amostragem e muito mais) que foram originalmente desenvolvidos no LinkedIn.	
Pydoop	O Pydoop é uma API MapReduce e HDFS do Python para Hadoop, criada com base nas APIs C ++ Pipes e C libhdfs, que permite criar aplicativos MapReduce completos com acesso HDFS. O Pydoop tem várias vantagens sobre as soluções integradas do Hadoop para programação em Python, ou seja, Hadoop Streaming e Jython: sendo um pacote CPython, ele permite que você acesse todos os módulos de bibliotecas padrão e de terceiros, alguns dos quais podem não estar disponíveis.	1. Site do SF Pydoop 2. Projeto Pydoop GitHub
Canguru	Projeto de código aberto do Conductor para escrever trabalhos MapReduce que consomem dados do Kafka. A postagem introdutória explica o caso de uso do Conductor - carregar dados do Kafka para o HBase por meio de um trabalho MapReduce usando o HFileOutputFormat. Ao contrário de outras soluções que são limitadas a uma única partição InputSplit por Kafka, o Kangaroo pode lançar vários consumidores em diferentes deslocamentos no fluxo de uma única partição para aumentar o throughput e o paralelismo.	1. Introdução ao canguru 2. Projeto Kangaroo GitHub
TinkerPop	Estrutura de computação gráfica escrita em Java. Fornece uma API principal que os fornecedores de sistemas gráficos podem implementar. Existem vários tipos de sistemas de gráficos, incluindo bibliotecas de gráficos na memória, bancos de dados de gráficos OLTP e processadores de gráficos OLAP. Uma vez que as interfaces centrais são implementadas, o sistema de gráficos subjacente pode ser consultado usando a linguagem gráfica traversal Gremlin e processado com algoritmos ativados pelo TinkerPop. Para muitos, o TinkerPop é visto como o JDBC da comunidade de computação gráfica.	1. Proposta do Apache Tinkerpop 2. Site do TinkerPop
Paquiderma MapReduce	O Pachyderm é um mecanismo MapReduce completamente novo, construído nos principais Docker e CoreOS. No Pachyderm MapReduce (PMR), um trabalho é um servidor HTTP dentro de um contêiner do Docker (um microsserviço). Você dá à Pachyderm	1. Local de paquiderme 2. Artigo de introdução de paquiderme

	uma imagem do Docker e a distribuirá automaticamente por todo o cluster ao lado de seus dados. Os dados são colocados no container por HTTP e os resultados são armazenados novamente no sistema de arquivos. Você pode implementar o servidor da Web em qualquer idioma desejado e acessar qualquer biblioteca. O Pachyderm também cria um DAG para todos os trabalhos no sistema e suas dependências e agenda automaticamente o pipeline de forma que cada trabalho não seja executado até que as dependências sejam concluídas. Tudo no Pachyderm “fala em diffs” para que ele saiba exatamente quais dados foram alterados e quais subconjuntos do pipeline precisam ser reexecutados. O CoreOS é um sistema operacional leve de código aberto baseado no Chrome OS, na verdade, o CoreOS é uma bifurcação do Chrome OS. O CoreOS fornece apenas a funcionalidade mínima necessária para implantar aplicativos dentro de contêineres de software, juntamente com mecanismos incorporados para descoberta de serviço e compartilhamento de configuração	
Feixe Apache	O Apache Beam é um modelo unificado de código-fonte aberto para definir e executar pipelines de processamento de dados paralelos, bem como um conjunto de SDKs específicos de linguagem para a construção de pipelines e Runners específicos de tempo de execução para executá-los. O modelo por trás da Beam evoluiu a partir de vários projetos internos de processamento de dados do Google, incluindo MapReduce, FlumeJava e Millwheel. Esse modelo era originalmente conhecido como "Modelo de fluxo de dados" e implementado pela primeira vez como Google Cloud Dataflow, incluindo um Java SDK no GitHub para escrever pipelines e serviço totalmente gerenciado para executá-los no Google Cloud Platform. Em janeiro de 2016, o Google e vários parceiros enviaram o modelo de programação do Dataflow e a parte dos SDKs como uma proposta da Apache Incubator, sob o nome de Apache Beam (processamento unificado Batch + strEAM).	1. Proposta do Apache Beam 2. Comparação de Feixes e Spark do DataFlow
	Bancos de dados NoSQL	
	Modelo de dados de coluna	
Apache HBase	Google BigTable Inspirado. Banco de dados distribuído não relacional. Ramdom, operações de r / w em tempo real em tabelas muito grandes orientadas por colunas	1. Apache HBase Home 2. Espelho do HBase no Github

	(BDDB: Big Data Data Base). É o sistema de apoio para saídas de trabalhos de MR. É o banco de dados do Hadoop. É para fazer backup de tarefas do Hadoop MapReduce com tabelas do Apache HBase	
Apache Cassandra	DBMS não-SQL distribuído, é um BDDB. O MR pode recuperar dados do Cassandra. Este BDDB pode ser executado sem HDFS ou no topo do HDFS (DataStax fork do Cassandra). O HBase e seus sistemas de suporte necessários são derivados do que é conhecido dos designs originais do Google BigTable e do Sistema de arquivos do Google (conforme conhecido do documento do Google File System publicado em 2003 e do BigTable publicado em 2006). Cassandra, por outro lado, é uma bifurcação recente de fonte aberta de um sistema de banco de dados autônomo inicialmente codificado pelo Facebook, que ao implementar o modelo de dados BigTable, usa um sistema inspirado pelo Dynamo da Amazon para armazenar dados (na verdade muito do trabalho de desenvolvimento inicial em Cassandra). foi realizado por dois engenheiros do Dynamo recrutados para o Facebook da Amazon).	1. Apache HBase Home 2. Cassandra no GitHub 3. Recursos de treinamento 4. Cassandra - Paper
Hypertable	Sistema de banco de dados inspirado em publicações sobre o design do BigTable do Google. O projeto é baseado na experiência de engenheiros que estavam resolvendo tarefas intensivas de dados em larga escala por muitos anos. O Hypertable é executado sobre um sistema de arquivos distribuído, como o Apache Hadoop DFS, o GlusterFS ou o Kosmos File System (KFS). Está escrito quase inteiramente em C ++. Sposored pelo Baidu, o mecanismo de busca chinês.	FAÇAM
Apache Accumulo	O armazenamento de chave / valor distribuído é um sistema de armazenamento e recuperação de dados robusto, escalável e de alto desempenho. O Apache Accumulo é baseado no design BigTable do Google e foi desenvolvido com base no Apache Hadoop, Zookeeper e Thrift. O Accumulo é um software criado pela NSA com recursos de segurança.	1. Apache Accumulo Home
Apache Kudu	Armazenamento de dados relacional, colunar, distribuído e otimizado para casos de uso analítico que exigem leituras muito rápidas com velocidades de gravação competitivas. • Modelo de dados relacionais (tabelas) com colunas fortemente tipadas e uma operação rápida e on-line de alteração de tabelas. • Expandir e dividir com suporte para particionamento com base em intervalos de chaves e / ou hashing.	1. Apache Kudu Home 2. Kudu no Github 3. Kudu whitepaper técnico (pdf)

	• Resistente a falhas e consistente devido à sua implementação do consenso de jangadas. • Suportado pelo Apache Impala e Apache Drill, permitindo rápidas leituras e gravações SQL através desses sistemas. • Integra-se com MapReduce e Spark. • Além disso, fornece APIs "NoSQL" em Java, Python e C ++.	
Parquet Apache	Formato de armazenamento colunar disponível para qualquer projeto no ecossistema Hadoop, independentemente da escolha da estrutura de processamento de dados, modelo de dados ou linguagem de programação.	1. Apache Parquet Home 2. Apache Parquet no Github
	Modelo de dados do documento	
MongoDB	Sistema de banco de dados orientado a documentos. Faz parte da família NoSQL de sistemas de banco de dados. Em vez de armazenar dados em tabelas, como é feito em um banco de dados relacional "clássico", o MongoDB armazena dados estruturados como documentos semelhantes a JSON	1. Site do Mongodb
RethinkDB	O RethinkDB é construído para armazenar documentos JSON e escalar para várias máquinas com muito pouco esforço. Ele tem uma linguagem de consulta agradável que suporta consultas realmente úteis, como junções de tabelas e agrupamento, e é fácil de configurar e aprender.	1. Site RethinkDB
ArangoDB	Um banco de dados de código aberto com um modelo de dados flexível para documentos, gráficos e valores-chave. Crie aplicativos de alto desempenho usando uma linguagem de consulta semelhante a um sql ou extensões JavaScript.	1. site ArangoDB
	Modelo de dados de fluxo	
EventStore	Um banco de dados funcional de código aberto com suporte para Processamento de Eventos Complexos. Ele fornece um mecanismo de persistência para aplicativos que usam a terceirização de eventos ou para armazenar dados de séries temporais. O Armazenamento de Eventos é escrito em C #, C ++ para o servidor que é executado no Mono ou no .NET CLR, no Linux ou no Windows. Aplicativos usando o Event Store podem ser escritos em JavaScript. Event sourcing (ES) é uma maneira de persistir o estado do seu aplicativo, armazenando o histórico que determina o estado atual do seu aplicativo.	1. Site do EventStore
	Modelo de dados de valor-chave	
Base de Dados Redis	O Redis é um armazenamento de estruturas de dados	1. Site do Redis 2.

	em memória aberta, de código aberto, com durabilidade opcional. Está escrito em ANSI C. Em sua camada externa, o modelo de dados Redis é um dicionário que mapeia chaves para valores. Uma das principais diferenças entre o Redis e outros sistemas de armazenamento estruturado é que o Redis suporta não apenas strings, mas também tipos de dados abstratos. Patrocinado pela Redis Labs. É licenciado pelo BSD.	Site do Redis Labs
Linkedin Voldemort	Armazenamento de dados distribuído que é projetado como um armazenamento de valor-chave usado pelo LinkedIn para armazenamento de alta escalabilidade.	1. site Voldemort
RocksDB	O RocksDB é um armazenamento persistente de valor-chave que pode ser incorporado para armazenamento rápido. RocksDB também pode ser a base para um banco de dados cliente-servidor, mas nosso foco atual é em cargas de trabalho incorporadas.	1. site do RocksDB
OpenTSDB	O OpenTSDB é um banco de dados de séries temporais (TSDB) distribuído e escalável, escrito em cima do HBase. O OpenTSDB foi escrito para atender a uma necessidade comum: armazenar, indexar e veicular métricas coletadas de sistemas de computadores (equipamentos de rede, sistemas operacionais, aplicativos) em larga escala e tornar esses dados facilmente acessíveis e grafitáveis.	1. site do OpenTSDB
	Modelo de Dados Gráficos	
ArangoDB	Um banco de dados de código aberto com um modelo de dados flexível para documentos, gráficos e valores-chave. Crie aplicativos de alto desempenho usando uma linguagem de consulta semelhante a um sql ou extensões JavaScript.	1. site ArangoDB
Neo4j	Um banco de dados de gráficos de código aberto escrito inteiramente em Java. É um mecanismo de persistência Java totalmente transacional e baseado em disco, que armazena dados estruturados em gráficos e não em tabelas.	1. site Neo4j
TitanDB	O TitanDB é um banco de dados de gráficos altamente escalonável, otimizado para armazenar e consultar grandes gráficos com bilhões de vértices e arestas distribuídas em um cluster de múltiplas máquinas. O Titan é um banco de dados transacional que pode suportar milhares de usuários simultâneos.	1. site do Titan
	Bancos de Dados NewSQL	
TokuDB	O TokuDB é um mecanismo de armazenamento para MySQL e MariaDB projetado especificamente para alto desempenho em cargas de trabalho intensivas em gravação. Isso é alcançado através da indexação da	1. site Percona TokuDB

	Árvore Fractal. O TokuDB é um mecanismo de armazenamento escalonável, compatível com ACID e MVCC. O TokuDB é uma das tecnologias que permitem Big Data no MySQL.	
HandlerSocket	O HandlerSocket é um plugin NoSQL para o MySQL / MariaDB (o mecanismo de armazenamento do MySQL). Ele funciona como um daemon dentro do processo mysqld, aceitando conexões TCP e executando solicitações de clientes. O HandlerSocket não suporta consultas SQL. Em vez disso, suporta operações CRUD simples em tabelas. O HandlerSocket pode ser muito mais rápido que o mysqld / libmysql em alguns casos, porque tem menor sobrecarga de CPU, disco e rede.	FAÇAM
Servidor Akiban	O Akiban Server é um banco de dados de código aberto que reúne os armazenamentos de documentos e os bancos de dados relacionais. Os desenvolvedores obtêm um poderoso acesso a documentos junto com um SQL surpreendentemente poderoso.	FAÇAM
Chuvisco	Drizzle é uma versão reprojetada da base de código MySQL v6.0 e é projetada em torno de um conceito central de ter uma arquitetura microkernel. Recursos como cache de consulta e sistema de autenticação são agora plugins para o banco de dados, que seguem o tema geral de "mecanismos de armazenamento plugáveis" que foram introduzidos no MySQL 5.1. Ele suporta PAM, LDAP e HTTP AUTH para autenticação através de plugins fornecidos. Por meio de seu sistema de plug-ins, ele atualmente suporta o registro em log de arquivos, syslog e serviços remotos, como RabbitMQ e Gearman. O Drizzle é um banco de dados relacional compatível com ACID que suporta transações através de um projeto de MVCC	FAÇAM
Haeinsa	Haeinsa é uma biblioteca de transações multi-linha e multi-table linearmente escalável para o HBase. Use Haeinsa se você precisar de semântica ACID forte em seu cluster HBase. É baseado no conceito do Google Perlocator.	1. Site do Haeinsa GitHub
SenseiDB	Banco de dados de código aberto, distribuído, em tempo real e semi-estruturado. Alguns recursos: pesquisa de texto completo, atualizações rápidas em tempo real, pesquisa estruturada e facetada, BQL: linguagem de consulta semelhante a SQL, pesquisa rápida de valor-chave, alto desempenho em volumes de consulta e atualização pesados simultâneos, integração do Hadoop	1. SenseiDB site
Céu	O Sky é um banco de dados de código aberto usado	1. site do SkyDB

	para análise flexível e de alto desempenho de dados comportamentais. Para determinados tipos de dados, como dados de fluxo de cliques e dados de log, ele pode ser várias ordens de magnitude mais rápido do que as abordagens tradicionais, como bancos de dados SQL ou Hadoop.	
BayesDB	O BayesDB, uma tabela de banco de dados bayesiana, permite aos usuários consultar as prováveis implicações de seus dados tabulares com a mesma facilidade com que um banco de dados SQL permite que eles consultem os dados em si. Usando a Linguagem de Consulta Bayesiana (BQL) integrada, os usuários sem treinamento em estatística podem resolver problemas básicos de ciência de dados, como detectar relações preditivas entre variáveis, inferir valores ausentes, simular observações prováveis e identificar entradas de banco de dados estatisticamente semelhantes.	1. site BayesDB
InfluxDB	O InfluxDB é um banco de dados de séries temporais distribuídas de software livre sem dependências externas. É útil para gravar métricas, eventos e realizar análises. Ele tem uma API HTTP embutida para que você não precise escrever nenhum código do lado do servidor para ser instalado e executado. O InfluxDB foi projetado para ser dimensionável, simples de instalar e gerenciar e rápido para obter dados. Destina-se a responder a consultas em tempo real. Isso significa que cada ponto de dados é indexado à medida que entra e está imediatamente disponível em consultas que devem retornar abaixo de 100 ms.	1. Site do InfluxDB
	SQL-on-Hadoop	
Apache Hive	Infraestrutura de Data Warehouse desenvolvida pelo Facebook. Sumarização, consulta e análise de dados. Ele fornece linguagem semelhante a SQL (não compatível com SQL92): HiveQL.	1. Site do Apache HIVE 2. Projeto Apache HIVE GitHub
Apache HCatalog	A abstração de tabelas do HCatalog apresenta aos usuários uma visão relacional de dados no HDFS (Hadoop Distributed File System) e garante que os usuários não precisem se preocupar sobre onde ou em que formato seus dados são armazenados. Neste momento, o HCatalog faz parte do Hive. Apenas versões antigas são separadas para download.	FAÇAM
Apache Trafodion	O Apache Trafodion é uma solução SQL-on-Hadoop de escala da Web que permite cargas de trabalho transacionais e operacionais de classe empresarial no HBase. O Trafodion é um mecanismo de banco de dados MPEG ANSI SQL nativo que se baseia na	1. Site do Apache Trafodion 2. Wiki do Trafodion Apache 3. Projeto GitHub do Trafodion Apache

	escalabilidade, elasticidade e flexibilidade do HDFS e HBase, estendendo-os para fornecer integridade transacional garantida para todas as cargas de trabalho, incluindo multi-coluna, várias linhas, várias tabelas e vários servidores atualizações.	
Apache HAWQ	O Apache HAWQ é um mecanismo de consulta SQL nativa do Hadoop que combina as principais vantagens tecnológicas do banco de dados MPP desenvolvido a partir do banco de dados Greenplum, com a escalabilidade e conveniência do Hadoop.	1. Site do Apache HAWQ 2. Projeto HAWQ GitHub
Apache Drill	O Drill é a versão de código aberto do sistema Dremel, do Google, que está disponível como um serviço de infraestrutura chamado Google BigQuery. Nos últimos anos, os sistemas de código aberto surgiram para atender à necessidade de processamento em lote escalável (Apache Hadoop) e processamento de fluxo (Storm, Apache S4). O Apache Hadoop, originalmente inspirado pelo sistema MapReduce interno do Google, é usado por milhares de organizações que processam conjuntos de dados em larga escala. O Apache Hadoop foi projetado para atingir um throughput muito alto, mas não foi projetado para atingir a latência de subsegundos necessária para análise e exploração interativa de dados. Drill, inspirado no sistema Dremel interno do Google, destina-se a atender a essa necessidade	1. Broca de Incubadora Apache
Cloudera Impala	O projeto Impala, licenciado pela Apache, traz a tecnologia de banco de dados paralelo escalável para o Hadoop, permitindo que os usuários emitam consultas SQL de baixa latência para dados armazenados no HDFS e no Apache HBase, sem exigir movimento ou transformação de dados. É um clone do Google Dremel (Big Query google).	1. Cloudera Impala site 2. Projeto Impala GitHub
Presto do Facebook	O Facebook criou o Presto, um mecanismo de SQL que, em média, é 10 vezes mais rápido que o Hive para executar consultas em grandes conjuntos de dados armazenados no Hadoop e em outros locais.	1. site do Presto
SQL de Splat de Datasalt	O Splout permite servir um conjunto de dados arbitrariamente grande com altas taxas de QPS e, ao mesmo tempo, fornece uma sintaxe de consulta SQL completa.	FAÇAM
Apache Tajo	O Apache Tajo é um robusto sistema de data warehouse distribuído e relacional de big data para o Apache Hadoop. O Tajo é projetado para consultas ad hoc escaláveis e de baixa latência, agregação on-line e ETL (processo de transferência de carga extraído) em grandes conjuntos de dados armazenados no HDFS	1. Site do Apache Tajo

	(Hadoop Distributed File System) e em outras fontes de dados. Suportando padrões SQL e aproveitando técnicas avançadas de banco de dados, o Tajo permite o controle direto da execução distribuída e do fluxo de dados em uma variedade de estratégias de avaliação de consultas e oportunidades de otimização. Para referência, a Apache Software Foundation anunciou o Tajo como um projeto de nível superior em abril de 2014.	
Apache Phoenix	O Apache Phoenix é um skin SQL sobre HBase fornecido como um driver JDBC incorporado ao cliente, que direciona consultas de baixa latência sobre dados do HBase. O Apache Phoenix pega sua consulta SQL, compila-a em uma série de varreduras do HBase e orquestra a execução dessas varreduras para produzir conjuntos de resultados regulares do JDBC. Os metadados da tabela são armazenados em uma tabela do HBase e versionados, de modo que as consultas de instantâneos sobre as versões anteriores usarão automaticamente o esquema correto. O uso direto da HBase API, juntamente com coprocessadores e filtros personalizados, resulta em desempenho na ordem de milissegundos para pequenas consultas ou segundos para dezenas de milhões de linhas.	1. Site do Apache Phoenix
Apache MRQL	O MRQL é um sistema de processamento e otimização de consultas para análise de dados distribuídos em grande escala, criado com base no Apache Hadoop, Hama e Spark. O MRQL (pronunciado milagre) é um sistema de processamento e otimização de consultas para análise de dados distribuídos em larga escala. O MRQL (a Linguagem de Consulta MapReduce) é uma linguagem de consulta semelhante a SQL para análise de dados em grande escala em um cluster de computadores. O sistema de processamento de consultas MRQL pode avaliar consultas MRQL em três modos: • no modo Map-Reduce usando o Apache Hadoop, • no modo BSP (Bulk Synchronous Parallel mode) usando o Apache Hama, e • no modo Spark usando o Apache Spark. • no modo Flink usando o Apache Flink.	1. Site MRQL da Incubadora Apache
Kylin	O Kylin é um mecanismo de análise distribuída de código aberto da eBay Inc. que fornece interface SQL e análise multidimensional (OLAP) no Hadoop, suportando conjuntos de dados extremamente grandes	1. site do projeto Kylin
	Ingestão de dados	

Apache Flume	O Flume é um serviço distribuído, confiável e disponível para coletar, agregar e mover com eficiência grandes quantidades de dados de log. Tem uma arquitetura simples e flexível baseada em fluxos de dados de fluxo contínuo. Ele é robusto e tolerante a falhas, com mecanismos de confiabilidade ajustáveis e muitos mecanismos de failover e recuperação. Ele usa um modelo de dados extensível simples que permite a aplicação analítica on-line.	1. Site do projeto Apache Flume
Apache Sqoop	Sistema para transferência de dados em massa entre HDFS e datastores estruturados como RDBMS. Como o Flume, mas do HDFS para o RDBMS.	1. Site do projeto Apache Sqoop
Escriba do Facebook	Agregador de log em tempo real. É um serviço Apache Thrift.	1. Site do Facebook Scribe GitHub
Apache Chukwa	Agregador de logs em grande escala e análises.	1. site Apache Chukwa
Apache Kafka	Sistema distribuído de publicação / assinatura para processamento de grandes quantidades de dados de fluxo. Kafka é uma fila de mensagens desenvolvida pelo LinkedIn que persiste mensagens no disco de uma maneira muito eficiente. Como as mensagens são persistentes, tem a capacidade interessante de os clientes retrocederem um fluxo e consumirem as mensagens novamente. Outra vantagem da persistência do disco é que a importação em massa dos dados para o HDFS para análise offline pode ser feita de forma muito rápida e eficiente. O Storm, desenvolvido pela BackType (que foi adquirida pelo Twitter há um ano), trata mais de transformar um fluxo de mensagens em novos fluxos.	1. Apache Kafka 2. Código-fonte do GitHub
Netflix Suro	O Suro tem suas raízes no Apache Chukwa, que foi inicialmente adotado pela Netflix. É um agregador de logs como Storm, Samza.	FAÇAM
Apache Samza	O Apache Samza é uma estrutura de processamento de fluxo distribuído. Ele usa o Apache Kafka para mensagens e o Apache Hadoop YARN para fornecer tolerância a falhas, isolamento do processador, segurança e gerenciamento de recursos. Desenvolvido por http://www.linkedin.com/in/jaykreps Linkedin.	1. Site do Apache Samza
Morphline de Cloudera	O Cloudera Morphlines é uma nova estrutura de código aberto que reduz o tempo e as habilidades necessárias para integrar, criar e alterar aplicativos de processamento do Hadoop que extraem, transformam e carregam dados no Apache Solr, Apache HBase, HDFS, data warehouses corporativos ou painéis on-line analíticos .	FAÇAM
HIHO	Este projeto é uma estrutura para conectar origens de	FAÇAM

	dados diferentes com o sistema Apache Hadoop, tornando-as interoperáveis. O HIHO conecta o Hadoop a vários sistemas RDBMS e de arquivos, para que os dados possam ser carregados no Hadoop e descarregados do Hadoop	
Apache NiFi	O Apache NiFi é um sistema de fluxo de dados que está atualmente em incubação na Apache Software Foundation. NiFi é baseado nos conceitos de programação baseada em fluxo e é altamente configurável. O NiFi usa um modelo de extensão baseado em componentes para adicionar recursos rapidamente a fluxos de dados complexos. Fora da caixa, a NiFi tem várias extensões para lidar com fluxos de dados baseados em arquivos, como FTP, SFTP e integração HTTP, bem como a integração com o HDFS. Um dos recursos exclusivos do NiFi é uma interface rica e baseada na Web para projetar, controlar e monitorar um fluxo de dados.	1. Apache NiFi
Manifold ApacheCF	O Apache ManifoldCF fornece uma estrutura para conectar repositórios de conteúdo de origem, como sistemas de arquivos, banco de dados, CMIS, SharePoint, FileNet ..., para repositórios de destino ou índices, como o Apache Solr ou o ElasticSearch. É um tipo de rastreador para repositórios com vários conteúdos, suportando muitas fontes e conversão de vários formatos para indexação por meio do filtro de transformação Apache Tika Content Extractor.	1. Manifold ApacheCF
	Programação de Serviço	
Apache Thrift	Uma estrutura de RPC entre linguagens para criações de serviço. É a base de serviços para as tecnologias do Facebook (o colaborador original do Thrift). O Thrift fornece uma estrutura para desenvolver e acessar serviços remotos. Ele permite que os desenvolvedores criem serviços que podem ser consumidos por qualquer aplicativo que esteja gravado em um idioma para o qual há ligações do Thrift. O Thrift gerencia a serialização de dados para e de um serviço, bem como o protocolo que descreve uma chamada de método, resposta, etc. Em vez de escrever todo o código RPC - você pode ir direto para a lógica de serviço. O Thrift usa o TCP e, portanto, um determinado serviço é ligado a uma porta específica.	1. Apache Thrift
Apache Zookeeper	É um serviço de coordenação que fornece as ferramentas necessárias para escrever aplicativos distribuídos corretos. O ZooKeeper foi desenvolvido no Yahoo! Pesquisa. Vários projetos do Hadoop já estão usando o ZooKeeper para coordenar o cluster e	1. Apache Zookeeper 2. Google Chubby paper

	fornecer serviços distribuídos altamente disponíveis. Talvez os mais famosos sejam Apache HBase, Storm, Kafka. O ZooKeeper é uma biblioteca de aplicativos com duas implementações principais das APIs - Java e C - e um componente de serviço implementado em Java que é executado em um conjunto de servidores dedicados. O Zookeeper é para construir sistemas distribuídos, simplifica o processo de desenvolvimento, tornando-o mais ágil e permitindo implementações mais robustas. Em 2006, o Google publicou um artigo sobre "Chubby", um serviço de trava distribuída que ganhou ampla adoção em seus data centers. Zookeeper,	
Apache Avro	O Apache Avro é uma estrutura para modelagem, serialização e criação de chamadas de procedimento remoto (RPC). Os dados do Avro são descritos por um esquema, e um recurso interessante é que o esquema é armazenado no mesmo arquivo que os dados que ele descreve, portanto, os arquivos são autoexplicativos. O Avro não requer geração de código. Essa estrutura pode competir com outras ferramentas semelhantes, como: Apache Thrift, Google Protocol Buffers, ZeroC ICE e assim por diante.	1. Apache Avro
Curador do Apache	Curador é um conjunto de bibliotecas Java que torna o uso do Apache ZooKeeper muito mais fácil.	FAÇAM
Apache karaf	O Apache Karaf é um tempo de execução do OSGi que é executado sobre qualquer estrutura do OSGi e fornece um conjunto de serviços, um poderoso conceito de provisionamento, um shell extensível e muito mais.	FAÇAM
Twitter elefante pássaro	O Elephant Bird é um projeto que fornece utilitários (bibliotecas) para trabalhar com dados compactados com LZOP. Ele também fornece um formato de contêiner que suporta o trabalho com Protocol Buffers, Thrift no MapReduce, Writables, Pig LoadFuncs, Hive SerDe, HBase miscellanea. Esta biblioteca de código aberto é usada maciçamente no Twitter.	1. GitHub de Pássaro Elefante
Linkedin Norbert	O Norbert é uma biblioteca que fornece fácil gerenciamento de cluster e distribuição de carga de trabalho. Com o Norbert, você pode distribuir rapidamente uma arquitetura cliente / servidor simples para criar uma arquitetura altamente escalável capaz de lidar com tráfego pesado. Implementado no Scala, Norbert envolve o ZooKeeper, Netty e usa Protocol Buffers para transporte, facilitando a criação de um aplicativo que reconhece clusters. Uma API Java é fornecida e estratégias de balanceamento de carga	1. Projeto Linkedin 2. Código-fonte do GitHub

	plugáveis são suportadas com round robin e estratégias de hash consistentes fornecidas prontas para uso.	
	Agendamento e DR	
Apache Oozie	Sistema do planejador de fluxo de trabalho para tarefas de RM usando DAGs (Direct Acyclical Graphs). Oozie Coordinator pode acionar jobs por tempo (frequência) e disponibilidade de dados	1. Apache Oozie 2. Código-fonte do GitHub
LinkedIn Azkaban	Gerenciamento de fluxo de trabalho do Hadoop. Um agendador de tarefas em lote pode ser visto como uma combinação do cron e tornar os utilitários Unix combinados com uma interface amigável.	LinkedIn Azkaban
Apache Falcon	O Apache Falcon é uma estrutura de gerenciamento de dados para simplificar o gerenciamento do ciclo de vida de dados e o processamento de pipelines no Apache Hadoop. Ele permite que os usuários configurem, gerenciem e orquestrem a movimentação de dados, o processamento de pipeline, a recuperação de desastres e os fluxos de trabalho de retenção de dados. Em vez de codificar intensamente os recursos de ciclo de vida de dados, os aplicativos do Hadoop agora podem contar com a estrutura do Apache Falcon bem testada para essas funções. A simplificação do gerenciamento de dados do Falcon é bastante útil para qualquer um que cria aplicativos no Hadoop. O gerenciamento de dados no Hadoop engloba movimentação de dados, orquestração de processos, gerenciamento do ciclo de vida, descoberta de dados, entre outras preocupações que estão além do ETL.	Apache Falcon
Horóscopo	O Schedoscope é um novo projeto de código aberto que fornece uma estrutura de agendamento para desenvolvimento, teste, (re) carregamento e monitoramento ágil de seu datahub, lago ou o que você escolher para chamar seu data warehouse do Hadoop atualmente. Conjuntos de dados (incluindo dependências) são definidos usando um scala DSL, que pode incorporar tarefas MapReduce, scripts Pig, consultas Hive ou fluxos de trabalho Oozie para criar o conjunto de dados. A ferramenta inclui uma estrutura de teste para verificar a lógica e um utilitário de linha de comando para carregar e recarregar dados.	Código-fonte do GitHub
	Aprendizado de Máquina	
Apache Mahout	Biblioteca de aprendizado de máquina e biblioteca de matemática, além de MapReduce.	Apache Mahout
WEKA	O Weka (Waikato Environment for Knowledge Analysis) é um conjunto popular de software de aprendizado de máquina escrito em Java, desenvolvido	Weka 3

	na Universidade de Waikato, Nova Zelândia. Weka é um software livre disponível sob a Licença Pública Geral GNU.	
Cloudera Oryx	O projeto de código aberto da Oryx fornece infra-estrutura simples e em tempo real de aprendizagem de máquina / análise preditiva em larga escala. Ele implementa algumas classes de algoritmos comumente usados em aplicativos de negócios: filtragem / recomendação colaborativa, classificação / regressão e clustering.	1. Oryx no GitHub 2. Fórum Cloudera para Machine Learning
Deeplearning4j	O projeto de código aberto Deeplearning4j é o framework de aprendizagem profunda mais utilizado para a JVM. O DL4J inclui redes neurais profundas, como redes neurais recorrentes, redes de longo prazo de memória (LSTMs), redes neurais convolucionais, vários autoencodificadores e redes neurais progressivas, como máquinas restritas de Boltzmann e redes de crenças profundas. Ele também possui algoritmos de processamento de linguagem naturais, como word2vec, doc2vec, GloVe e TF-IDF. Todas as redes Deeplearning4j são executadas distribuídas em várias CPUs e GPUs. Eles funcionam como jobs do Hadoop e integram-se ao Spark no nível slace para orquestração de threads de host. As redes neurais do Deeplearning4j são aplicadas a casos como casos de fraude e detecção de anomalias, sistemas de recomendação e manutenção preditiva.	1. Deeplearning4j Website 2. Comunidade Gitter para o Deeplearning4j
MADlib	O projeto MADlib utiliza os recursos de processamento de dados de um RDBMS para analisar dados. O objetivo deste projeto é a integração da análise de dados estatísticos em bancos de dados. O projeto MADlib é auto-descrito como o Big Data Machine Learning em SQL para Data Scientists. O projeto de software MADlib começou no ano seguinte como uma colaboração entre pesquisadores da UC Berkeley e engenheiros e cientistas de dados da EMC / Greenplum (agora Pivotal)	1. Comunidade MADlib
H2O	H2O é uma ferramenta estatística, de aprendizado de máquina e de matemática para análise de bigdata. Desenvolvida pela empresa de análise preditiva H2O.ai, a H2O estabeleceu uma liderança na cena ML juntamente com a Spark da R e Databricks. De acordo com a equipe, a H2O é a plataforma de memória mais rápida do mundo para aprendizado de máquina e análise preditiva em big data. Ele é projetado para ajudar os usuários a dimensionar o aprendizado de máquina, a matemática e as estatísticas em grandes conjuntos de dados.	1. H2O no GitHub 2. H2O Blog

	Além do ponto de H2O e clique em Web-UI, sua API REST permite fácil integração em vários clientes. Isso significa que a análise exploratória de dados pode ser feita de maneira típica em R, Python e Scala; e fluxos de trabalho inteiros podem ser gravados como scripts automatizados.	
Água com gás	O Sparkling Water combina duas tecnologias de código aberto: Apache Spark e H2O - um mecanismo de aprendizado de máquina. Ele faz a biblioteca de algoritmos avançados do H2O, incluindo Aprendizado Profundo, GLM, GBM, KMeans, PCA e Random Forest, acessíveis a partir dos fluxos de trabalho Spark. Os usuários do Spark são fornecidos com as opções para selecionar os melhores recursos das plataformas para atender às necessidades de Aprendizado de Máquina. Os usuários podem combinar a API RDD do Sparks e o Spark MLLib com os algoritmos de aprendizado de máquina do H2O, ou usar o H2O independente do Spark no processo de criação do modelo e pós-processar os resultados no Spark. O Sparkling Water fornece uma integração transparente do framework e das estruturas de dados do H2O no ambiente baseado em RDD do Spark compartilhando o mesmo espaço de execução, além de fornecer uma API semelhante ao RDD para estruturas de dados H2O.	1. Água com gás no GitHub 2. Exemplos de água com gás
Apache SystemML	O Apache SystemML foi aberto pela IBM e está bastante relacionado com o Apache Spark. Se você está pensando no Apache Spark como o sistema operacional de análise para qualquer aplicativo que aproveite grandes volumes de dados de streaming. O MLLib, a biblioteca de aprendizado de máquina do Spark, fornece aos desenvolvedores um rico conjunto de algoritmos de aprendizado de máquina. E o SystemML permite que os desenvolvedores traduzam esses algoritmos para que possam digerir facilmente diferentes tipos de dados e executá-los em diferentes tipos de computadores. O SystemML permite que um desenvolvedor grave um algoritmo de aprendizado de máquina único e dimensione-o automaticamente usando o Spark ou o Hadoop. O SystemML é escalável para análise de big data com tecnologia otimizadora de alto desempenho e permite que os usuários escrevam algoritmos personalizados de aprendizado de máquina usando linguagem simples e específica de domínio (DSL) sem aprender	1. Apache SystemML 2. Proposta do Apache

	programação distribuída complicada. É um framework de complementação extensível do Spark MLlib.	
	Ferramentas de Benchmarking e QA	
Benchmarking do Apache Hadoop	Existem dois arquivos JAR principais no Apache Hadoop para benchmarking. Este JAR são micro-benchmarks para testar partes específicas da infraestrutura, por exemplo TestDFSIO analisa o sistema de disco, TeraSort avalia tarefas MapReduce, o WordCount mede o desempenho do cluster, etc. Micro-Benchmarks são empacotados nos arquivos JAR de testes e exmaples, e você pode obtenha uma lista deles, com descrições, chamando o arquivo JAR sem argumentos. Com relação à versão estável do Apache Hadoop 2.2.0, temos disponíveis os seguintes arquivos JAR para teste, exemplos e benchmarking. Os micro-benchmarks do Hadoop são empacotados nesses arquivos JAR: hadoop-mapreduce-examples-2.2.0.jar, hadoop-mapreduce-client-jobclient-2.2.0-tests.jar.	1. MAPREDUCE-3561 guarda-chuva bilhete para rastrear todas as questões relacionadas com o desempenho
Yahoo Gridmix3	Hadoop cluster benchmarking da equipe de engenheiros do Yahoo.	FAÇAM
Benchmarking PUMA	Conjunto de referência que representa uma ampla variedade de aplicativos MapReduce exibindo características de aplicação com computação de alta / baixa e volumes de embaralhamento altos / baixos. Há um total de 13 benchmarks, dos quais Tera-Sort, Word-Count e Grep são da distribuição do Hadoop. Os demais benchmarks foram desenvolvidos internamente e atualmente não fazem parte da distribuição do Hadoop. Os três benchmarks da distribuição do Hadoop também são levemente modificados para obter o número de tarefas de redução como entrada do usuário e gerar estatísticas finais de conclusão de tarefas.	1. MAPREDUCE-5116 2. Faraz Ahmad pesquisador 3. PUMA Docs
Berkeley SWIM Benchmark	O benchmark SWIM (Statistical Workload Injector para MapReduce), é uma referência que representa uma carga de trabalho de big data do mundo real desenvolvida pela Universidade da Califórnia em Berkley, em estreita cooperação com o Facebook. Esse teste fornece medições rigorosas do desempenho de sistemas MapReduce compostos de cargas de trabalho reais da indústria.	1. GitHub SWIN
Intel HiBench	O HiBench é uma suíte de benchmark do Hadoop.	FAÇAM
Apache Yetus	Para ajudar a manter a consistência em um conjunto grande e desconectado de committers, o teste de patch automatizado foi adicionado ao processo de desenvolvimento do Hadoop. Esse teste de patch	1. Entrada do blog de alta qualidade 2. Proposta do Apache Yetus 3. Site

	automatizado (agora incluído como parte do Apache Yetus) funciona da seguinte maneira: quando um patch é carregado no sistema de acompanhamento de bugs, um processo automatizado faz o download do patch, realiza algumas análises estáticas e executa os testes de unidade. Esses resultados são postados de volta no rastreador de bugs e alertas notificam as partes interessadas sobre o estado do patch. No entanto, o projeto Apache Yetus aborda muito mais do que o tradicional teste de patch, é uma abordagem melhor, incluindo uma reescrita massiva do recurso de teste de patch usado no Hadoop.	do Projeto Apache Yetus
	Segurança	
Sentinela Apache	O Sentry é a próxima etapa em segurança de Big Data de nível corporativo e fornece uma autorização refinada aos dados armazenados no Apache Hadoop. Um módulo de segurança independente que se integra com os mecanismos de consulta SQL de código-fonte aberto Apache Hive e Cloudera Impala, o Sentry oferece controles avançados de autorização para permitir aplicativos multiusuários e processos multifuncionais para conjuntos de dados corporativos. Sentinela era um desenvolvimento de Cloudera.	FAÇAM
Portal do Apache Knox	Sistema que fornece um único ponto de acesso seguro para clusters do Apache Hadoop. O objetivo é simplificar a segurança do Hadoop para os usuários (ou seja, quem acessa os dados do cluster e executar tarefas) e os operadores (ou seja, quem controla o acesso e gerencia o cluster). O Gateway é executado como um servidor (ou cluster de servidores) que atende a um ou mais clusters do Hadoop.	1. Apache Knox 2. Portal Apache Knox Hortonworks web
Patrulheiro Apache	O Apache Argus Ranger (anteriormente chamado de Apache Argus ou HDP Advanced Security) oferece uma abordagem abrangente para a administração central de políticas de segurança nos principais requisitos corporativos de segurança de autenticação, autorização, contabilidade e proteção de dados. Ele estende os recursos de linha de base para aplicação coordenada em cargas de trabalho do Hadoop de SQL e tempo real em lote e interativo e aproveita a arquitetura extensível para aplicar políticas consistentemente em componentes do ecossistema Hadoop adicionais (além de HDFS, Hive e HBase), incluindo Storm, Solr, Spark e Mais.	1. Apache Ranger 2. Apache Ranger Hortonworks web
	Gerenciamento de metadados	
Metacoscópio	O Metascope é uma ferramenta de gerenciamento de	Código-fonte do

	metadados e descoberta de dados que serve como um complemento para o Schedoscope. O Metascope é capaz de coletar metadados técnicos, operacionais e comerciais de seu Hadoop Datahub e facilita a busca e a navegação por meio de um portal.	GitHub
	Implantação do sistema	
Apache Ambari	Interface web de gerenciamento intuitiva e fácil de usar do Hadoop, apoiada por suas APIs RESTful. O Apache Ambari foi doado pela equipe da Hortonworks para o ASF. É uma interface poderosa e agradável para o Hadoop e outros aplicativos típicos do ecossistema Hadoop. O Apache Ambari está sob um desenvolvimento pesado e incorporará novos recursos em um futuro próximo. Por exemplo, o Ambari é capaz de implantar um sistema Hadoop completo a partir do zero, no entanto, não é possível usar essa GUI em um sistema Hadoop que já está em execução. A capacidade de provisionar o sistema operacional pode ser uma boa adição, mas provavelmente não está no roteiro.	1. Apache Ambari
Cloudera HUE	Aplicativo da Web para interagir com o Apache Hadoop. Não é uma ferramenta de deploment, é uma interface da Web de código aberto que suporta o Apache Hadoop e seu ecossistema, licenciado sob a licença Apache v2. O HUE é usado para o Hadoop e suas operações do usuário do ecossistema. Por exemplo, a HUE oferece editores para Hive, Impala, Oozie, Pig, cadernos para Spark, painéis de pesquisa do Solr, HDFS, YARN, navegadores do HBase.	1. Página inicial do HUE
Apache Mesos	Mesos é um gerenciador de cluster que fornece compartilhamento de recursos e isolamento entre aplicativos de cluster. Tal como a HTCondor, a SGE ou a Troque podem fazê-lo. No entanto Mesos é design centrado no hadoop	FAÇAM
Miríade	Myriad é uma estrutura mesos projetada para escalar clusters YARN no Mesos. O Myriad pode expandir ou reduzir um ou mais clusters YARN em resposta a eventos conforme regras e políticas configuradas.	1. Myriad Github
Maratona	Marathon é uma estrutura Mesos para serviços de longa duração. Dado que você tem o Mesos rodando como o kernel para o seu datacenter, o Marathon é o daemon init ou upstart.	FAÇAM
Brooklyn	O Brooklyn é uma biblioteca que simplifica a implantação e o gerenciamento de aplicativos. Para a implantação, ele foi projetado para se conectar a outras ferramentas, oferecendo implementação com um único clique e adicionando os conceitos de clusters e malhas gerenciáveis: Muitas entidades comuns de software	FAÇAM

	disponíveis para uso imediato. Integra-se com o Apache Whirr - e, portanto, Chef e Puppet - para implantar serviços bem conhecidos, como Hadoop e elasticsearch (ou usar POBS, plain-old-bash-scripts) Use PaaS como o OpenShift, juntamente com clusters máxima flexibilidade	
Hortonworks HOYA	HOYA é definido como “executando o HBase On YARN”. A ferramenta Hoya é uma ferramenta Java e atualmente é acionada pelo CLI. Ele aceita uma especificação de cluster - em termos do número de servidores de regiões, o local de HBASE_HOME, os hosts de quorum do ZooKeeper, a configuração que a nova instância de cluster do HBase deve usar e assim por diante. Então HOYA é para implantação do HBase usando uma ferramenta desenvolvida em cima do YARN. Depois que o cluster for iniciado, o cluster poderá aumentar ou diminuir usando os comandos Hoya. O cluster também pode ser interrompido e depois retomado. A Hoya implementa a funcionalidade por meio das APIs do YARN e dos shell scripts do HBase. O objetivo do protótipo era ter alterações mínimas no código e, desde a sua elaboração, exigiu mudanças de código zero no HBase.	1. Hortonworks Blog
Apache Helix	O Apache Helix é uma estrutura genérica de gerenciamento de cluster usada para o gerenciamento automático de recursos particionados, replicados e distribuídos hospedados em um cluster de nós. Originalmente desenvolvido pelo Linkedin, agora está em um projeto de incubadora no Apache. O Helix é desenvolvido no topo do Zookeeper para tarefas de coordenação.	1. Apache Helix
Apache Bigtop	O Bigtop foi originalmente desenvolvido e lançado como uma infraestrutura de pacotes de código aberto pela Cloudera. O BigTop é usado por alguns fornecedores para construir suas próprias distribuições baseadas no Apache Hadoop (CDH, Pivotal HD, distribuição da Intel), no entanto o Apache Bigtop faz muito mais tarefas, como teste de integração contínua (com Jenkins, maven, ...) e é útil para empacotamento (RPM e DEB), implantação com o Puppet e assim por diante. O BigTop também apresenta receitas vagabundas para criar clusters hadoop "n-node" e o aplicativo blueprint bigpetstore, que demonstra a construção de um aplicativo hadoop de pilha completa com ETL, aprendizado de máquina e geração de conjunto de dados. O Apache Bigtop poderia ser considerado como um esforço da comunidade com um	1. Apache Bigtop.

	foco principal: colocar todos os bits do ecossistema do Hadoop como um todo, em vez de projetos individuais.	
Buildoop	O Buildoop é um projeto de código aberto licenciado sob o Apache License 2.0, baseado na ideia do Apache BigTop. O Buildoop é um projeto de colaboração que fornece modelos e ferramentas para ajudá-lo a criar sistemas personalizados baseados no Linux com base no ecossistema Hadoop. O projeto é construído a partir do scrach usando a linguagem Groovy, e não é baseado em uma mistura de ferramentas como o BigTop (Makefile, Gradle, Groovy, Maven), provavelmente é mais fácil de programar do que o BigTop, e o design é focado nas idéias básicas por trás o projeto Yocto buildroot. O projeto está em estágios iniciais de desenvolvimento agora.	1. Construtor de Ecossistema Hadoop.
Deploop	O Deploop é uma ferramenta para provisionar, gerenciar e monitorar clusters do Apache Hadoop focados na Arquitetura Lambda. LA é um projeto genérico baseado nos conceitos do engenheiro do Twitter Nathan Marz. Essa arquitetura genérica foi projetada para atender a requisitos comuns de big data. O sistema Deploop está em desenvolvimento contínuo, em fases alfa de maturidade. O sistema está configurado com tecnologias altamente escaláveis, como o Puppet e o MCollective.	1. O Sistema de Implementação do Hadoop.
SequenceIQ Cloudbreak	O Cloudbreak é uma maneira eficiente de iniciar e executar várias instâncias e versões de clusters do Hadoop na nuvem, contêineres do Docker ou bare metal. É uma API de plataforma Hadoop As-a-Service agnóstica e com custo reduzido em nuvem e infraestrutura. Fornece dimensionamento automático, multilocação segura e gerenciamento completo do ciclo de vida da nuvem. O Cloudbreak aproveita as plataformas de infraestrutura em nuvem para criar instâncias de host, usa a tecnologia Docker para implantar os containers necessários na nuvem e usa o Apache Ambari (via Ambari Blueprints) para instalar e gerenciar um cluster Hortonworks. Esta é uma ferramenta dentro do ecossistema HDP.	1. Projeto GitHub. 2. Introdução Cloudbreak. 3. Cloudbreak em Hortonworks.
Águia Apache	O Apache Eagle é uma solução de análise de código aberto para identificar problemas de segurança e desempenho instantaneamente em plataformas de big data, por exemplo, Hadoop, Spark etc. Ele analisa atividades de dados, aplicativos de fio, métricas jmx e logs de daemon etc. mecanismo de alerta de arte para identificar violações de segurança, problemas de	1. Projeto Apache Eagle Github. 2. Site da Apache Eagle.

	desempenho e mostra insights. A plataforma de Big Data normalmente gera uma quantidade enorme de logs e métricas operacionais em tempo real. A Apache Eagle é fundada para resolver problemas difíceis na segurança e ajuste de desempenho para plataformas de big data, garantindo métricas, logs sempre disponíveis e alertando imediatamente, mesmo sob tráfego intenso.	
	Aplicações	
Apache Nutch	Projeto de software de rastreador da Web de software livre altamente extensível e escalável. Um mecanismo de pesquisa baseado no Lucene: Um rastreador da Web é um bot da Internet que navega sistematicamente na World Wide Web, geralmente para fins de indexação da Web. Os rastreadores da Web podem copiar todas as páginas visitadas para processamento posterior por um mecanismo de pesquisa que indexe as páginas baixadas para que os usuários possam pesquisá-las com muito mais rapidez.	FAÇAM
Servidor de pesquisa do Sphinx	O Sphinx permite que você indexe em lote e pesquise dados armazenados em um banco de dados SQL, armazenamento NoSQL ou apenas arquivos de maneira rápida e fácil - ou indexe e pesquise dados dinamicamente, trabalhando com o Sphinx como um servidor de banco de dados.	1. Sphinx search web site
Apache OODT	OODT foi originalmente desenvolvido no Laboratório de Propulsão a Jato da NASA para apoiar a captura, processamento e compartilhamento de dados para arquivos científicos da NASA.	FAÇAM
Biblioteca HIPI	O HIPI é uma biblioteca da estrutura MapReduce do Hadoop que fornece uma API para executar tarefas de processamento de imagem em um ambiente de computação distribuída.	FAÇAM
PivotalR	O PivotalR é um pacote que permite que os usuários de R, a mais conhecida linguagem de programação estatística e de código-fonte aberto, interajam com o banco de dados Pivotal (Greenplum), bem como com o Pivotal HD / HAWQ e com o banco de dados de código aberto PostgreSQL for Big Data analytics. R é uma linguagem de programação e software de análise de dados: você faz a análise de dados em R escrevendo scripts e funções na linguagem de programação R. R é uma linguagem completa, interativa e orientada a objetos: projetada por estatísticos, para estatísticos. A linguagem fornece objetos, operadores e funções que tornam o processo de exploração, modelagem e visualização de dados natural.	1. PivotalR no GitHub
	Frameworks de Desenvolvimento	

Jumbune	O Jumbune é um produto de código aberto que está no topo de qualquer distribuição do Hadoop e auxilia no desenvolvimento e administração de soluções MapReduce. O objetivo do produto é auxiliar fornecedores de soluções analíticas a portar aplicativos sem falhas em ambientes de produção do Hadoop. O Jumbune suporta todos os principais ramos ativos do Apache Hadoop, nomeadamente 1.x, 2.x, 0.23.xe distribuições comerciais MapR, HDP 2.xe CDH 5.x do Hadoop. Ele tem a capacidade de funcionar bem com as versões Yarn e não-Yarn do Hadoop. Ele possui quatro módulos principais, o MapReduce Debugger, o HDFS Data Validator, o monitor de cluster sob demanda e o mapeador de tarefas MapReduce. O Jumbune pode ser implementado em qualquer máquina remota do usuário e usa um agente leve no NameNode do cluster para transmitir informações relevantes para lá e para cá.	1. Jumbune 2. Projeto Jumbune GitHub 3. Jumbune JIRA page
Spring XD	O Spring XD (Xtreme Data) é uma evolução da estrutura de desenvolvimento de aplicativos Spring Java para ajudar aplicativos de Big Data da Pivotal. A SpringSource foi a empresa criada pelos fundadores do Spring Framework. A SpringSource foi comprada pela VMware, onde foi mantida por algum tempo como uma divisão separada dentro da VMware. Mais tarde, a VMware e sua empresa matriz, a EMC Corporation, criaram formalmente uma joint venture chamada Pivotal. O Spring XD é mais do que uma biblioteca de estruturas de desenvolvimento, é um sistema distribuído e extensível para ingestão de dados, análise em tempo real, processamento em lote e exportação de dados. Pode ser considerado como alternativa ao Apache Flume / Sqoop / Oozie em alguns cenários. O Spring XD faz parte do Pivotal Spring para o Apache Hadoop (SHDP). SHDP, integrado com Spring, Spring Batch e Spring Data fazem parte da Spring IO Platform como bibliotecas fundamentais. Com base nisso, e estendendo essa base, a plataforma Spring IO fornece o Spring XD como tempo de execução de big data. O Spring for Apache Hadoop (SHDP) visa ajudar a simplificar o desenvolvimento de aplicativos baseados no Hadoop fornecendo uma configuração consistente e API em uma ampla variedade de projetos do ecossistema Hadoop, como Pig, Hive e Cascading, além de fornecer extensões ao Spring Batch para orquestrar Fluxos de trabalho baseados no Hadoop.	1. Spring XD no GitHub
Plataforma de Aplicação de Dados em Cask	O Cask Data Application Platform é uma plataforma de desenvolvimento de aplicativos de código aberto para o	1. Site do Cask

	ecossistema Hadoop que fornece aos desenvolvedores dados e virtualização de aplicativos para acelerar o desenvolvimento de aplicativos, lidar com vários casos de uso em tempo real e em lotes e implantar aplicativos na produção. A implantação é feita pelo Cask Coopr, uma solução de gerenciamento de cluster baseada em modelo de software livre que provisiona, gerencia e dimensiona clusters para pilhas de aplicativos de várias camadas em nuvens públicas e privadas. Outro componente é o Tigon, uma estrutura distribuída baseada no Apache Hadoop e no Apache HBase para aplicativos de análise e processamento de dados em tempo real, com alta taxa de transferência e baixa latência.	
	Categorizar pendente ...	
Apache Fluo	O Apache Fluo (incubação) é uma implementação de código aberto do Percolator for Apache Accumulo. O Fluo possibilita atualizar incrementalmente os resultados de um cálculo, índice ou analítica em larga escala à medida que novos dados são descobertos. O Fluo permite processar novos dados com menor latência do que o Spark ou o Map Reduce, no caso em que todos os dados devem ser reprocessados quando novos dados chegam.	1. Apache Fluo Site 2. Percolator Paper
Twitter Summingbird	Um sistema que visa mitigar as compensações entre o processamento em lote e o processamento em fluxo, combinando-os em um sistema híbrido. No caso do Twitter, o Hadoop lida com o processamento em lote, o Storm lida com o processamento de fluxo e o sistema híbrido é chamado de Summingbird.	FAÇAM
Apache Kiji	Crie aplicativos de Big Data em tempo real no Apache HBase.	FAÇAM
S4 Yahoo	O S4 é uma plataforma plugável, tolerante a falhas, distribuída, distribuída e de uso geral que permite que os programadores desenvolvam facilmente aplicativos para o processamento contínuo de fluxos ilimitados de dados.	FAÇAM
Metamarkers Druid	Armazenamento de dados analíticos em tempo real.	FAÇAM
Cascata Concorrente	Framework de aplicativos para desenvolvedores Java para simplesmente desenvolver aplicativos robustos de Data Analytics e Data Management no Apache Hadoop.	FAÇAM
Lingual Concorrente	Projeto de código aberto que permite o desenvolvimento rápido e simples de aplicativos Big Data no Apache Hadoop. projeto que fornece	FAÇAM

	tecnologia SQL padrão da ANSI para criar facilmente novos aplicativos integrados ao Hadoop	
Padrão Concorrente	Aprendizado de máquina para conexão em cascata no Apache Hadoop por meio de uma API e PMML baseada em padrões	FAÇAM
Apache Giraph	O Apache Giraph é um sistema de processamento gráfico interativo desenvolvido para alta escalabilidade. Por exemplo, atualmente é usado no Facebook para analisar o gráfico social formado pelos usuários e suas conexões. Giraph se originou como a contraparte de código aberto da Pregel, a arquitetura de processamento de grafos desenvolvida no Google	FAÇAM
Talend	A Talend é uma fornecedora de software de código aberto que fornece integração de dados, gerenciamento de dados, integração de aplicativos corporativos e soluções e softwares de big data.	FAÇAM
Akka Toolkit	O Akka é um kit de ferramentas e um tempo de execução de código aberto que simplificam a construção de aplicativos simultâneos na plataforma Java.	FAÇAM
Eclipse BIRT	O BIRT é um sistema de geração de relatórios baseado em Eclipse de software livre que se integra ao seu aplicativo Java / Java EE para produzir relatórios atraentes.	FAÇAM
Spango BI	A SpagoBI é uma suíte de Open Source Business Intelligence, pertencente à iniciativa SpagoWorld gratuita / de código aberto, fundada e apoiada pelo Engineering Group. Ele oferece uma ampla gama de funções analíticas, uma camada semântica altamente funcional, muitas vezes ausente em outras plataformas e projetos de código aberto, e um conjunto respeitável de recursos avançados de visualização de dados, incluindo análises geoespaciais.	FAÇAM
Jedox Palo	O Palo Suite combina todos os principais aplicativos - OLAP Server, Palo Web, Palo ETL Server e Palo for Excel - em uma plataforma de Business Intelligence abrangente e personalizável. A plataforma é totalmente baseada em produtos de código aberto que representam uma solução de Business Intelligence de alto nível que está disponível totalmente livre de qualquer taxa de licença.	FAÇAM
Twitter Finagle	O Finagle é uma pilha de rede assíncrona para a JVM que você pode usar para construir clientes e servidores assíncronos de Chamada de Procedimento Remoto (RPC) em Java, Scala ou qualquer linguagem hospedada pela JVM.	FAÇAM

Intel GraphBuilder	Biblioteca que fornece ferramentas para construir gráficos de grande escala sobre o Apache Hadoop	FAÇAM
Apache Tika	O Toolkit detecta e extrai metadados e conteúdo de texto estruturado de vários documentos usando bibliotecas de analisador existentes.	FAÇAM
Apache Zeppelin	O Zeppelin é uma ferramenta moderna baseada na Web para os cientistas de dados colaborarem em projetos de exploração e visualização de dados em larga escala. É um interpretador de estilo de notebook que permite o compartilhamento de sessões de análise colaborativa entre usuários. O Zeppelin é independente da própria estrutura de execução. A versão atual é executada sobre o Apache Spark, mas possui APIs de intérprete conectáveis para suportar outros sistemas de processamento de dados. Mais estruturas de execução podem ser adicionadas em uma data posterior, como o Apache Flink, o Crunch, bem como back-ends do tipo SQL, como Hive, Tajo, MRQL.	1. Site do Apache Zeppelin
Névoa hidrosfera	O Hydrosphere Mist é um serviço para expor os trabalhos analíticos e os modelos de aprendizado de máquina do Apache Spark como serviços da Web em tempo real, em lotes ou reativos. Ele atua como um middleware entre o Apache Spark e a pilha de aprendizado de máquina e aplicativos voltados ao usuário.	1. github Hidrosphere Mist

Publicado com páginas do GitHub por Javi Roman e colaboradores