1. Hadoop został zainspirowany rozwiązaniami wymyślonymi przez:
    a. Cloudera
    b. Google
    c. Amazon
    d. Hortonworks
2. W kolekcji wynikowej operacji mapToPair
    a. jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej
    b. jest zawsze więcej elementów niż w kolekcji źródłowej
    c. może być mniej lub więcej elementów niż w kolekcji źródłowej, ale nigdy tyle samo
    d. liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej)
    e. jest zawsze mniej elementów niż w kolekcji źródłowej

    Komentarz: mapToPair(func)

    Mapuje elementy ze źródłowej RDD na elementy typu klucz-wartość. Każdy z elementów powstałej "dwójki" może mieć swój odrębny typ. Moim zdaniem mniej ale nie jestem pewien

3. Podaj rok pierwszego wydania Hadoop (wersja 0.1.0) – można się pomylić o 2 lata
    Odpowiedź: 2006
    Komentarz: April 2006, In March 2006, Owen O'Malley was the first committer to add to the Hadoop project; Hadoop 0.1.0 was released in April 2006. It continues to evolve through contributions that are being made to the project.
4. Ekspresyjny język zapytań w grafowych bazach danych, odpowiednik SQLa z baz relacyjnych to:
    a. GraphSQL
    b. GraphDB
    c. Cypher
    d. Cyber
5. NetLogo to
    a. System modelowania i symulacji wieloagentowych
    b. System grafowych baz danych
    c. Język programowania oparty na języku Logo

Komentarz: Wiki - NetLogo to język programowania i zintegrowane środowisko modelowania matematycznego. Eng wiki - NetLogo is a programming language and integrated development environment (IDE) for agent-based modeling

6. W procesie próbkowania Snowball sampling
    a. Węzły sąsiadujące są odwiedzane wielokrotnie
    b. Węzły sąsiadujące są odwiedzane tylko wtedy, gdy nie zostały odwiedzone w poprzednich iteracjach
    c. Węzły sąsiadujące są odwiedzane tylko wtedy gdy przerwa między iteracjami jest większa niż zadana wartość
7. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście HDFS
    a. Klienci zapisują kopie danych bezpośrednio do każdego DataNode
    b. Klienci zapisują dane do NameNode
    c. W przypadku zatrzymania się węzła NameNode, klaster staje się niedostępny

   d. **Jest zoptymalizowany do dużych, strumieniowych odczytów plików**
   e. Nie ma możliwości modyfikacji plików, można tylko dopisać dane na końcu pliku.
8. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście Ranger
   a. Wspiera kontrolę dostępu opartą na atrybutach
   b. Wykorzystywany w Hadoop od wersji 1.0
   c. Wspiera kontrolę dostępu opartą na rolach
   d. Umożliwia pracę jednego klastra dla wiele różnych firm (model chmurowy)

Komentarz: Enhanced support for different authorization methods - Role based access control, attribute based access control etc.

9. Podstawowy komponent Spark to
   a. Spark RDD
   b. Spark Main
   c. Spark Base
   d. Spark Core
10. W kolekcji wynikowej operacji flatMap
   a. liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej)
   b. jest zawsze więcej elementów niż w kolekcji źródłowej
   c. jest zawsze mniej elementów niż w kolekcji źródłowej
   d. jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej
   e. może być mniej lub więcej elementów niż w kolekcji źródłowej, ale nigdy tyle samo
11. W komponencie Spark SQL
   a. Import danych jest możliwy z dowolnego systemu baz danych, dla którego istnieje sterownik typu JDBC
   b. Import danych z bazy jest możliwy tylko poprzez plik SQL (dump file)
   c. Import danych jest ograniczony do kilku obsługiwanych przez Spark systemów baz danych

Komentarz: Import danych z systemu zarządzania bazą danych z wykorzystaniem interfejsu JDBC. Wymagane jest dołączenie do projektu sterownika JDBC dla używanej bazy danych

12. Formaty plików wykorzystywane w HIVE to:
   a. ORC
   b. tekstowy
   c. Parquet
   d. Avro
   e. sekwencyjny
13. Domyślna liczba utrzymywanych w HDFS kopii danych to (podaj liczbę)
   Odpowiedź: 2
14. W komponencie Spark GraphX podstawową strukturą danych jest
   a. Skierowany multigraf
   b. Graf skierowany
   c. Graf nieskierowany
15. Zaznacz właściwie nazwy typów operacji wykonywanych na obiektach RDD w Spark
   a. Transformacje
   b. Akcje
   c. Transakcje
   d. Funkcje lambda

16. Próbkowanie sieci oparte na algorytmie Snowball sampling bazuje na:
    a. Włączeniu do próbki tylko jednego sąsiada przetwarzanego węzła
    b. Włączeniu do próbki n sąsiadów przetwarzanego węzła
    c. Włączeniu do próbki n losowo wybranych węzłów z sieci
17. W procesie próbkowania Random Walk
    a. Możliwe są przeskoki do innych segmentów sieci
    b. Nie są możliwe przeskoki do innych segmentów sieci
    c. Przeskoki do innych segmentów sieci umożliwiają pozyskiwanie próbek z wyizolowanych obszarów do których nie można dotrzeć z innych węzłów
18. Próbkowanie sieci oparte na algorytmie Random Walk bazuje na
    a. Włączeniu do próbki n losowo wybranych węzłów z sieci
    b. Włączeniu do próbki na sąsiadów przetwarzanego węzła
    c. Włączaniu do próbki tylko jednego losowo wybranego sąsiada przetwarzanego węzła
19. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście TEZ
    a. Przyśpiesza obliczenia
    b. Współdziała tylko z HIVE
    c. Wykorzystuje acykliczny graf skierowany do ustalenia kolejności zadań
    d. Pośrednie wyniki obliczeń zapisywane są do HDFS

Komentarz: Przyspiesza obliczenia -> "Silnik wykonujący lepiej operacje MapReduce (ok. 10x szybciej)" "DAG - (Directed- Acyclic-Graph) – skierowany graf acykliczny" Reszty nie wiem

20. Hadoop działa w trybie
    a. schema-on-read
    b. muszą być zdefiniowane transformacje danych nim dane będą załadowane
    c. schema-on-write
    d. nowe kolumny muszą być zdefiniowane nim dane będą załadowane
21. Algorytmami szeregowania w YARN są
    a. FIFO
    b. Capacity
    c. Priority
    d. LIFO
    e. Fair
22. Dopasuj hasło do opisu w kontekście Zookeepera:
    Klient zobaczy ten sam widok systemu, niezależnie od serwera, z którym się łączy. (Pojedynczy obraz systemu)
    Aktualizacje od poszczególnych klientów są stosowane w kolejności ich wysyłania. (Spójność sekwencyjna)
    Serwer z nieświeżymi danych zostanie zamknięty, zmuszając klienta do przełączenia się na bardziej aktualny serwer. (Aktualność)
    Aktualizacje przetrwają awarie serwerów. (Trwałość)
    Aktualizacje albo się udają albo nie. (Atomowość)
23. Język SQL trudno stosować do analizy struktur sieciowych ponieważ:
    a. działa zbyt wolno
    b. są one zbyt duże
    c. nie posiada dedykowanych analiz sieciowych
24. Dataset API w porównaniu do RDD API zapewnia
    a. Interfejs programistyczny wykorzystujący język zapytań SQL

b. Większe możliwości importu danych z różnych źródeł
c. Większą wydajność obliczeniową

25. Zaznacz nazwy istniejących komponentów Spark
    a. Spark Streaming
    b. Spark Graphics
    c. Spark Structural Data
    d. Spark SQL
    e. Spark ALS
    f. Spark GraphX
26. Apache Spark to (wybierz najbardziej pasującą odpowiedź)
    a. Biblioteka programistyczna wspomagająca budowę systemów rekomendujących
    b. Biblioteka programistyczna zawierająca metody uczenia maszynowego
    c. Platforma przyspieszająca dostęp do dużych zbiorów danych
    d. Platforma programistyczna do obliczeń rozproszonych
27. W komponencie Spark Streaming strumień danych przychodzących reprezentowany jest przez typ
    a. DatasetStream
    b. RDDStream
    c. RDD
    d. DStream
28. RDD API zapewnia taką samą wydajność obliczeń bez względu na zastosowany język programowania
    a. Prawda
    b. Fałsz

Komentarz – Jak jest wykres wydajności dla języków, to python stoi w tyle…

29. Zaznacz zadania realizowane przez Zookeepera
    a. Wykrywanie awarii jednego z serwerów (robotników lub nadzorcy)
    b. Wybór serwera, który będzie pełnił rolę nadzorcy (mastera)
    c. Wyłączanie węzłów klastra, które nie są chwilowo potrzebne z powodu mniejszego obciążenia zadaniami.
    d. Zarządzanie metadanymi przechowywanymi w znode'ach
    e. Zarządzanie plikami przechowywanymi w HDFS
30. Do metod próbkowania eksploracyjnego sieci zaliczamy
    a. Snow ball sampling
    b. Random Walk
    c. Random Edge selection
31. W języku Cypher odpowiednikiem polecenia SELECT jest
    a. PATH
    b. SEARCH
    c. MATCH
32. Platforma Apache Kafka wymaga przesyłania komunikatów w formacie:
    a. nie wymaga żadnego formatu, może on być dowolny
    b. XML
    c. JSON
    d. AVRO
33. W komponencie Spark GraphX graf można utworzyć z
    a. Samej kolekcji krawędzi

b. Kolekcji krawędzi kolekcji wierzchołków
c. Samej kolekcji wierzchołków

34. Domyślny rozmiar bloku danych w HDFS to
    a. 256 MB
    b. 256 kB
    c. 8 MB
    d. 32 kB
    e. 32 MB
    f. 128 MB
    g. 128 kB
    h. 8 kB
35. Apache Spark do prawidłowego działania wymaga dostępu do platformy Hadoop
    a. Prawda
    b. Fałsz
36. W komponencie Spark SQL zapytania w języku SQL są możliwe
    a. Dla każdej kolekcji Dataset bez względu na źródło danych
    b. Wyłącznie dla kolekcji Dataset powstałych poprzez zaimportowanie danych z relacyjnej bazy danych
37. W komponencie Spark GraphX wartości jakie można przypisać do krawędzi i węzłów grafu to
    a. Wyłącznie wartości typów prostych (numeryczne, tekstowe, logiczne)
    b. Dowolne wartości (obiekty lub wartości typów prostych)
    c. Wyłącznie wartości liczbowe
38. Komunikaty przesyłane w Apache Kafka
    a. mogą mieć wiele źródeł (producentów danych)
    b. mogą mieć wyłącznie jedno źródło (producenta danych)
    c. mogą mieć wyłącznie jednego odbiorcę (konsumenta danych)
    d. mogą mieć wielu odbiorców (konsumentów danych)
39. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście Phoenix
    a. Obsługuje transakcje
    b. Umożliwia analizę danych w trybie OLTP
    c. W pełni obsługuję operację JOIN
    d. wspiera i wykorzystuje pomocnicze indeksy (secondary indexes)
    e. Umożliwia dynamiczną rozbudowę schematu kolumn w czasie pracy.
40. Wynikiem zapytań w grafowej bazie danych Neo4j może być
    a. Ścieżka łącząca zadane węzły
    b. Wizualizacja grafu
    c. Tabela z numerycznymi i tekstowymi danymi wynikowymi
41. Dopasuj opis do nazwy kategorii:
    Protokół bezpiecznego uwierzytelniania: **Kerberos**
    Przechowuje polityki bezpieczeństwa: **Ranger**
    NoSQL'owa baza danych dla Hadoop: **HBase**
    Umożliwia SQL dostęp do danych w Hadoop: **Phoenix**
    Firewall dla hadoop: **Knox**
    Zarządza pracą klastra: **ZooKeeper**
    Interfejs graficzny zarządzania klastrem: **Ambari**
42. Przewaga w wydajności Spark względem Hadoop MapReduce wynika głównie z:
    a. Bardziej efektywnych algorytmów i modeli reprezentacji danych

b. Zdefiniowania w specyfikacji Spark standardów określających wymagania techniczne klastrów obliczeniowych
c. Lepszego wykorzystania pamięci RAM maszyn w klastrze obliczeniowym
d. Zastosowania innego języka programowania

Komentarz:

"Jest rozwinięciem idei z Hadoop MapReduce zapewniającym większą wydajność i większe możliwości analityczne"

"Jego bardzo duża szybkość wynika z wykorzystania głównie pamięci RAM w przeciwieństwie do bazującego na zapisach na dysku Hadoop MapReduce"

"Przewaga w wydajności wynika także z zastosowanych metod reprezentacji danych w silniku obliczeniowym bazujących na skierowanych grafach acyklicznych (DAG Engine)"

43. W modelu rozprzestrzeniania SIR
    a. Węzły sieci po wyzdrowieniu mogą ponownie się zarazić
    b. Węzły sieci po wyzdrowieniu mogą nadal infekować
    c. Węzły sieci po wyzdrowieniu nie biorą udziału w procesie propagacji

    Komentarz: SIR = Susceptible, Infectious, or Recovere

44. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście HIVE?
    a. Dane tabel są przechowywane w katalogu w HDFS
    b. Jest w pełni zgodny z ACID
    c. Dane są przechowywane w Metastore (???)
    d. Wykorzystywany jest mechanizm optymalizacji zapytań

    Komentarz: Optymalizacja -> internet + grafy z prezentacji (istnieje optymalizator w sterowniku). Tabela jest przechowywana w katalogu w HDFS. Meta dane są przechowywane w metastore (nwm czy dane) CHEATS ;-;

45. Wskaż cechy hadoop
    a. ma 100% zgodność z SQL
    b. działa w trybie OLAP
    c. działa w trybie ACID
    d. ma elastyczną strukturę danych
    e. umożliwia złożone przetwarzanie danych
    f. jest skalowalny
46. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście YARN?
    a. Pozwala na pracę wielu silników przetwarzających dane na jednym klastrze
    b. Zarządca zasobów działa w węzłach podrzędnych (DataNode)
    c. Zadania odpalane są kontenerach
    d. Jest jeden Application Master dla wszystkich zapytań
    e. Jest warstwą przetwarzania Hadoopa
47. Które języki programowania posiadają oficjalne wsparcie dla Apache Spark
    a. Perl
    b. Swift

c. **Java**
d. Kotlin
e. **Scala**
f. **Python**

48. Sqoop umożliwia:
    a. **Tranfer danych do HBase**
    b. **Transfer danych między Hadoop, a wieloma bazami relacyjnymi**
    c. Transfer danych tylko między Hadoop i mysql
49. Dataset API zapewnia taką samą wydajność obliczeń bez względu na zastosowany język programowania:
    a. Prawda
    b. Fałsz

    Komentarz: W wykladzie spark SQL gdzie jest dataset: "Przetwarzanie jest niezależne od użytego API czy języka programowania"

50. Które założenia bezpieczeństwa w Hadoop są prawdziwe?
    a. **Komunikacja HDFS i MapReduce nie będzie działać w niezaufanych sieciach**
    b. Hadoop tworzy konta dla użytkowników
    c. **Domyślnie zadanie w klastrze może trwać maksymalnie 7 dni**
    d. Użytkownicy muszą mieć dostęp do kont root w klastrze
    e. **Dostęp do HDFS będzie autoryzowany za pomocą tokenów**

    Komentarz: Wykład -> "Komunikacja HDFS i MapReduce nie będzie działać w niezaufanych sieciach"-> "Zadanie Hadoop będzie działać nie dłużej niż 7 dni (konfigurowalne) w klastrze MapReduce lub dostęp do HDFS z zadania zakończy się niepowodzeniem"-> Bilety Kerberos nie będą przechowywane w zadaniach MapReduce i nie będą dostępne dla zadań zadania. Dostęp do HDFS będzie autoryzowany za pomocą tokenów

51. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście MapReduce?
    a. **Wartości związane z danym kluczem zawsze trafią do tego samego reduktora**
    b. Najpierw odbywa się faza redukcji a następnie mapowania
    c. **Maper odczytuje dane w postaci par klucz/wartość**
    d. **Hadoop stara się zapewnić, że mapery działają na węzłach, które przechowują lokalnie swoją część danych**
    e. Wynik pracy węzła jest na bieżąco wysyłany do kolejnego węzła
52. Język SQL trudno stosować do analizy struktur sieciowych ponieważ:
    a. działa zbyt wolno
    b. **są one zbyt duże**
    c. **nie posiada dedykowanych analiz sieciowych**
53. Rozprzestrzenianie informacji w mediach społecznościowych może być modelowane z wykorzystaniem:
    a. **modeli epidemiologicznych**
    b. **modelu SIR**
    c. **systemów agentowych**
54. Podstawą do wyznaczenia miary closeness w analizach sieci są:
    a. najdłuższe ścieżki
    b. **najkrótsze ścieżki**
    c. połączenia dwukierunkowe
55. Stopień wierzchołka:

a. **jest jedną z miar centralności sieci**
b. jest podstawą do wyznaczania długości ścieżek
c. **określa liczbę sąsiadów węzła**

56. Możliwe stany węzłów sieci w modelu rozprzestrzeniania informacji SIR to:
    a. Selected Infected Recovered
    b. Susceptible Interested Recovered
    c. **Susceptible Infected Recovered**
57. Grafowa baza danych:
    a. **zapewnia dedykowane funkcje przetwarzania grafów**
    b. **zapewnia możliwość wyznaczania najkrótszych ścieżek**
    c. wprowadza nowe funkcje do języka SQL
58. Miara pośrednictwa w strukturach sieciowych określa:
    a. **Znaczenie węzła w przesyle informacji pomiędzy segmentami sieci**
    b. liczbę kontaktów w transmisji informacji
    c. Liczbę sąsiadów pośredniczących w przesyle informacji

1.Co oznacza szybkość w big data:

Dane powstają i są dostarczane niezwykle szybko, muszą być obsługiwane z odpowiednim reżimem czasowym. Obsługa olbrzymich ilości danych w czasie rzeczywistym

2. Co oznacza złożoność w big data:

Złożoność (complexity) - dane napływają z różnych źródeł, wymaga to łączenia, dopasowywania, oczyszczania danych

3. Co się składa na efektywność Hadoopa
    - szybki zapis
    - przeprowadza obliczenia tam, gdzie znajdują się dane
    - duże, strumieniowe odczyty plików
4. Co oznacza schema-on-read
    - dotyczy Hadoopa, NIE relacyjnych baz danych (relacyjne mają schema-on-write)
    - dane są kopiowane do pliku bez żadnej transformacji
    - mechanizm serializacji wykorzystany do odczytu danych (uzyskanie dostępu do kolumn)
    - szybki zapis
    - elastyczność
5. Wymienić 3 tryby YARNa
    - Demon NameNode - musi być bez przerwy uruchomiony, w przypadku zatrzymania się NameNode klaster przestanie być dostępny
    - tryb High - mamy dwa NameNodes: aktywny i w stanie gotowości
    - tryb klasyczny - główny NameNode i 1 węzeł pomocniczy
6. Na czym polega zadanie mapowania w MapReduce
    - polega na rozłożeniu zadań na wiele węzłów, działa na węźle, na którym przechowuje dane
    - maper odczytuje i zapisuje dane w postaci par klucz/wartość
7. Wymień komponenty Sparka
    - Spark MLLib (uczenie maszynowe, np ALS do systemów rekomendacyjnych)
    - GraphX
    - Spark SQL
    - Spark Streaming
    - GraphFrame (dodatkowy), nie jest oficjalną częścią sparka, wymaga dołączenia biblioteki
8. Wymień języki programowania dla Sparka
    - Java

- Scala
- Python
- R

9. Ocenianie jakości próbkowania
    - Próbkowanie można uznać za efektywne jeśli:
        - uzyskana próbka zachowuje właściwości sieci - w jaki sposób próbka aproksymuje własności całej sieci) - kryterium 1 - analizy własności próbki dają wyniki jak dla kompletnej sieci (centralność, ścieżki)
        - kryterium 2 - własności predykcyjne - czy model predykcyjny uczony z wykorzystaniem próbki umożliwi predykcję nieznanych właściwości
        - uzyskana próbka jest mniejsza niż sieć pierwotna
10. Dwie najważniejsze metody próbkowania sieci homogenicznych
    - wybór węzłów i krawędzi o zadanych właściwościach
    - pobieranie próbek w procesie eksploracji
        - Random Walk
        - Snowball sampling
        - poszukiwanie wzorców
11. 3 volume z bigdata
    - volume (objętość)
    - velocity (szybkość)
    - variety (różnorodność)
    - variability (zmienność)
12. Ekosystem Hadoopa 10 pozycji
    - ZooKeeper
    - hadoop MapReduce
    - Hive
    - Beeline
    - Sqoop
    - Hbase
    - Cassandra
    - Spark
        - Spark SQL
        - Spark MLLib (do systemów rekomendacyjnych używaliśmy)
        - Spark Streaming
        - GraphX
    - Pig
    - Strom
    - kafka
13. Filter w apache spark

    To funkcja będąca argumentem, zwracająca wartość logiczną (true lub false), decyduje, czy element zostanie przypisany do wynikowego RDD, służy do selekcji danych
14. Różnica między map a Flat Map

    map - wykonuje funkcję dla każdego elementu z pierwotnego RDD, wynik zapisuje w wynikowym RDD. Elementy w wynikowym RDD nie muszą być tego samego typu, co w pierwotnym, jednak zawsze jest ich tyle samo

    flatMap - wynikiem działania funkcji może być 0 lub kilka nowych elementów w nowym RDD, rozmiary pierwotnego i wynikowego elementu mogą być więc RÓŻNE
15. Lotnisko grafy czemu
16. Skąd pochodzą dane w big data
    - transakcje biznesowe
    - media społecznościowe

- dane z sensorów
- dane wymieniane pomiędzy urządzeniami
- dane strumieniowe
- dane z mediów społecznościowych
- dane dostępne publicznie
- dane z wewnętrznych baz danych

17. Różnica między RDD, DataSet, dataFrame

    DataFrame - kolekcje danych zorganizowane w kolumny, każda kolumna ma nazwę i typ

    DataSet - kolekcje, w których schemat danych zdefiniowany jest przez klasę JVM

    RDD - model rozproszonych danych

18. Hive

    Jest zbliżonym do SQL interfejsem wykorzystującym Hadoop MapReduce, koncentruje się na ekstrakcji, ładowaniu i przetwarzaniu:

    - odczytywanie ogromnych ilości danych
    - przekształcenia danych (mieszanie, konsolidacja, agregowanie)
    - załadowanie danych wyjściowych do innych systemów w celu dalszej analizy
    - klient CLI do Hive to: Beeline lub Hive

Big Data pytania egzamin wykład

1. Dopasuj opis do nazwy technologii
    a. Firewall dla Hadoop - Knox
    b. Interfejs graficzny do zarządzania klastrem - Ambari
    c. Zarządza pracą klastra – ZooKeeper
    d. Przechowuje polityki bezpieczeństwa - Ranger
    e. NoSQL'owa baza danych dla Hadoopa – HBase
    f. Umożliwia SQL dostęp do danych w Hadoop - Phoenix
    g. Protokół bezpiecznego uwierzytelniania - Kerberos
2. Próbkowanie sieci oparte na algorytmie Random Walk bazuje na:
    a. Włączaniu do próbki n sąsiadów przetwarzanego węzła
    b. Włączaniu do próbki tylko jednego losowo wybranego sąsiada przetwarzanego węzła
    c. Włączaniu do próbki n losowo wybranych węzłów sieci
3. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście MapReduce:
    a. Maper odczytuje dane w postaci par klucz/wartość
    b. Wynik pracy węzła jest na bieżąco wysyłany do kolejnego węzła
    c. Wartości związane z danym kluczem zawsze trafią do tego samego reduktora
    d. Najpierw odbywa się faza redukcji a następnie mapowania
    e. Hadoop stara się zapewnić, że mapery działają na węzłach, które przechowują lokalnie swoją część danych
4. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście HIVE:
    a. Jest w pełni zgodny z ACID
    b. Dane tabel są przechowywane w katalogu w HDFS
    c. Dane są przechowywane w Metastore
    d. Wykorzystywany jest mechanizm optymalizacji zapytań
5. Domyślny rozmiar bloku danych w HDFS to:
    a. 256 kB
    b. 256 MB
    c. 8 kB
    d. 32 MB
    e. 8 MB
    f. 128 MB
    g. 32 kB
    h. 128 kB
6. Wskaż cechy HADOOP:
    a. Umożliwia złożone przetwarzanie danych
    b. Działa w trybie OLAP
    c. Ma elastyczną strukturę danych
    d. Działa w trybie ACID
    e. Jest skalowalny
    f. Ma 100% zgodności z SQL
7. Które z poniższych stwierdzeń są prawdziwe w kontekście YARN:
    a. Zarządca zasobów działa w węzłach podrzędnych (DataNode)
    b. Pozwala na pracę wielu silników przetwarzających dane na jednym klastrze
    c. Jest jeden Application Master dla wszystkich zadań – napisane że jeden na apke
    d. Jest warstwą przetwarzania Hadoopa
    e. Zadania odpalane są w kontenerach – aplikacje wykonywane w 1/kilku kontenerach

8. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście TEZ:
    a. Pośrednie wyniki obliczeń zapisywane są do HDFS
    b. Wykorzystuje acykliczny graf skierowany do ustalenia kolejności zadań
    c. Współdziała tylko z HIVE
    d. Przyśpiesza obliczenia
9. Sqoop umożliwia:
    a. Transfer danych między Hadoop a wieloma bazami relacyjnymi
    b. Transfer danych tylko między Hadoop i mysql
    c. Transfer danych do HBase
10. Które założenia bezpieczeństwa w Hadoop są PRAWDZIWE:
    a. Domyślnie zadanie w klastrze może trwać maksymalnie 7 dni
    b. Komunikacja HDFS i MapReduce nie będzie działać w niezaufanych sieciach
    c. Hadoop tworzy konta dla użytkowników
    d. Użytkownicy musza mieć dostęp do kont root w klastrze
    e. Dostęp do HDFS będzie autoryzowany za pomocą tokenów – (logowanie uzyskuje token)
11. Które z poniższych stwierdzeń są PRAWDZIWE w kontekście Ranger:
    a. Wspiera kontrolę dostępu opartą na rolach
    b. Umożliwia pracę jednego klastra dla wielu różnych firm (model chmurowy)
    c. Wspiera kontrolę dostępu opartą na atrybutach
    d. Wykorzystywany w Hadoop od wersji 1.0
12. Ekspresyjny język zapytań w grafowych bazach danych, odpowiednik SQLa z baz relacyjnych to:
    a. GraphSQL
    b. Cypher – (z tego korzysta neo4j)
    c. Cyber
    d. GraphBD
13. NetLogo to:
    a. Język programowania oparty na języku Logo
    b. System modelowania i symulacji wieloagentowych
    c. System grafowych baz danych
14. W modelu rozprzestrzeniania SIR
    a. Węzły sieci po wyzdrowieniu mogą nadal infekować
    b. Węzły sieci po wyzdrowieniu mogą ponownie się zarazić
    c. Węzły sieci po wyzdrowieniu nie biorą udziału w procesie propagacji – (info o trwałej odporności)
15. Próbkowanie sieci oparte na algorytmie Snowball sampling bazuje na
    a. Włączaniu do próbki n sąsiadów przetwarzanego węzła
    b. Włączaniu do próbki tylko jednego sąsiada przetwarzanego sygnału
    c. Włączaniu do próbki n losowo wybranych węzłów z sieci
16. Które języki programowania posiadają oficjalne wsparcie dla Apache Spark?
    a. Perl
    b. Swift
    c. Kotlin
    d. Scala
    e. Python
17. W kolekcji wynikowej operacji flatMap
    a. Jest zawsze więcej elementów niż w kolekcji źródłowej

b. Jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej
c. Może być mniej lub więcej elementów niż w kolekcji źródłowej ale nigdy tyle samo
d. Jest zawsze mniej elementów niż w kolekcji źródłowej
e. Liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej) – (rozmiary RDD pierwotnego i wynikowego mogą być różne)

18. W kolekcji wynikowej operacji mapToPair:
    a. Jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej – (bo musi każdemu elementowi przypisać element do pary?)
    b. Może być mniej lub więcej elementów niż w kolekcji źródłowej ale nigdy tyle samo
    c. Jest zawsze mniej elementów niż w kolekcji źródłowej
    d. Liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej)
    e. Jest zawsze więcej elementów niż w kolekcji źródłowej
19. W komponencie Spark SQL
    a. Import danych z bazy jest możliwy tylko poprzez plik SQL (dump file)
    b. Import danych jest możliwy z dowolnego systemu bazy danych, dla którego istnieje sterownik typu JDBC
    c. Import danych jest ograniczony do kilku obsługiwanych przez Spark systemów baz danych
20. W komponencie Spark SQL zapytania w języku SQL są możliwe
    a. Wyłącznie dla kolekcji Dataset powstałych poprzez zaimportowanie danych z relacyjnej bazy danych
    b. Dla każdej kolekcji Dataset bez względu na źródło danych – (to wynika z tego że można wczytywać różne typy danych?)
21. Apache Spark to
    a. Platforma przyspieszająca dostęp do dużych zbiorów danych
    b. Biblioteka programistyczna wspomagająca budowę systemów rekomendujących
    c. Biblioteka programistyczna zawierająca metody uczenia maszynowego
    d. Platforma programistyczna do obliczeń rozproszonych
22. W komponencie Spark GraphX wartości jakie można przypisać do krawędzi i węzłów grafu to:
    a. Dowolne wartości (obiekty lub wartości typów prostych)
    b. Wyłącznie wartości typów prostych (numeryczne, tekstowe, logiczne)
    c. Wyłącznie wartości liczbowe
23. W komponencie Spark Streaming strumień danych przychodzących reprezentowany jest przez typ:
    a. DStream
    b. DataStream
    c. RDD
    d. RDDStream
24. Przewaga wydajności Spark względem Hadoop MapReduce wynika głównie z :
    a. Lepszego wykorzystania pamięci RAM maszyn w klastrze obliczeniowym
    b. Bardziej efektywnych algorytmach i modeli reprezentacji danych
    c. Zdefiniowania w specyfikacji Spark standardów określających wymagania techniczne dla klastrów obliczeniowych
    d. Zastosowania innego języka programowania
25. Podstawowy komponent Spark to:
    a. Spark Core
    b. Spark Main

c. Spark Base
d. Spark RDD

26. Zaznacz właściwe nazwy typów operacji wykonywanych na obiektach RDD w Spark:
    a. Transakcje
    b. Akcje
    c. Funkcje lambda
    d. Transformacje
27. Podaj rok pierwszego wydania Hadoop (wersja 0.1.0) – można się pomylić o 2 lata
    a. 2006 Kwiecień
28. RDD API zapewnia taką samą wydajność obliczeń bez względu na zastosowany język programowania:
    a. Prawda
    b. Fałsz – (tu coś mówił że w pythonie działa wolniej niż w Scali?)
29. Dataset API zapewnia taką samą wydajność obliczeń bez względu na zastosowany język programowania
    a. Prawda – (znalazłem info o tym że wszystkie języki używają tej samej optymalizacji kodu)
    b. Fałsz
30. Apache Spark do prawidłowego działania wymaga dostępu do platformy Hadoop
    a. Prawda
    b. Fałsz
31. Mechanizm szablonów zapytań SQL stosowany w interfejsach programistycznych baz danych (jedna odp):
    a. Umożliwia przesyłanie wartości parametrów do szablonu zapytania SQL zapisanego wcześniej w systemie bazy danych
    b. Umożliwia wstawienie wartości parametrów do zapytania SQL i następnie przesłanie kompletnego zapytania do systemu bazy danych w celu wykonania
32. Query Builder w interfejsach programistycznych bazy danych (wiele odp):
    a. Wprowadza alternatywny dla SQL język zapytań do bazy danych
    b. Pozwala na budowanie zapytań w oparciu o łańcuchowo wywoływane metody formułujące poszczególne ich części
    c. Zmniejsza ryzyko sformułowania niepoprawnego zapytania
    d. Umożliwia automatyczne generowanie struktury tabel w bazie danych na podstawie pól klas zdefiniowanych w aplikacji
33. W interfejsach programistycznych baz danych niemożliwe jest mapowanie klas zawierających pola o typach użytkownika (zdefiniowane klasy lub typy wyliczeniowe), możliwe jest wyłącznie mapowanie klas zawierając pola o typach prostych(liczbowe, tekstowe, numeryczne, logiczne)
    a. Prawda
    b. Fałsz
34. Menadżer encji w interfejsach programistycznych baz danych to (wiele odp):
    a. wzorzec projektowy klasy odpowiadającej za zapis i odczyt w bazie danych obiektów reprezentujących dane w aplikacji
    b. rola członka zespołu w projekcie informatycznym odpowiedzialna za projekt i implementację struktur bazy danych
    c. wzorzec projektowy klasy umożliwiający tworzenie nowych obiektów reprezentujących dane przetwarzane w aplikacji
35. Technika mapowania obiektowo-relacyjnego (wiele odp):

a. Wymaga stosowania w obiektowym modelu danych wyłącznie typów danych dostępnych w wykorzystywanym systemie bazodanowym
b. Automatyzuje proces zakładania, definiowania i aktualizowania struktur tabel w bazie danych
c. Umożliwia reprezentowanie struktur obiektowych w relacyjnej bazie danych
d. Podnosi wydajność operacji zapisu i odczytu bazy danych

36. Interfejsy programistyczne API działające warstwie abstrakcji w dostępie do danych (jedna odp):
    a. Wprowadzają abstrakcyjne struktury reprezentujące i modelujące dane (np. obiektowo)
    b. Umożliwiają automatyczne generowanie tabel w relacyjnej bazie danych
    c. Zapewniają metody jednolitej komunikacji z bazami danych od różnych dostawców
37. Sterowniki w interfejsach programistycznych dla relacyjnych baz danych (takich jak np. PDO lub JDBC) (jedna odp)
    a. umożliwiają rozszerzenie możliwości obsługiwanych baz danych o nowe typy danych dla kolumn oraz dodatkowe mechanizmy indeksowania
    b. zapewniają możliwość jednolitego komunikowania się z bazami danych od różnych dostawców
    c. umożliwiają rozszerzenie tych interfejsów o nowsze wersje języka zapytań SQL
38. Które z wymienionych interfejsów programistycznych działają w warstwie abstrakcji bazy danych (wiele odp) - to chyba cos innego niz abstrakcja W dostępie do danych (PDO)
    a. Java/Jakarta Persistence API (JPA)
    b. Doctrine ORM
    c. Java Database Conectivity (JDBC)
    d. Doctrine DBAL
    e. PHP Data Objects (PDO)
39. Kaskadowość w technikach mapowania obiektowo-relacyjnego polega na: (jedna odpowiedź)
    a. Wczytywaniu kolekcji obiektów jednego typu w kolejności zgodnej z wartościami identyfikatora,
    b. Możliwość łączenia operacji na bazie danych w grupy i wykonywania ich w ustalonej kolejności
    c. Wczytywaniu, zapisywaniu lub usuwaniu obiektu właściciela relacji wraz z obiektami podległymi
40. Czy w interfejsach programistycznych baz danych wykorzystujących techniki ORM możliwe jest mapowanie klas powstałych w wyniku dziedziczenia? (jedna odp)
    a. Prawda
    b. Fałsz
41. Pule połączeń (ang. connection pool) stosowane są w interfejsach programistycznych baz danych (wiele odp):
    a. Redukują konieczny narzut czasowy potrzebny na nawiązywanie nowego połączenia przez aplikację
    b. Ograniczają blokowanie dostępu do połączenia z bazą danych w aplikacjach wielowątkowych
    c. Zapewniają aplikacji dostęp do zestawu wielu jednocześnie otwartych połączeń do różnych baz danych
    d. Zapewniają aplikacji dostęp do zestawu wielu jednocześnie otwartych połączeń do tej samej bazy danych
42. Asocjacja jednokierunkowa w mapowaniu obiektowo-relacyjnym oznacza: (jedna odp)
    a. Relację pomiędzy obiektami, w której wyłącznie strona właściciela relacji posiada referencję do obiektu (lub obiektów) strony przeciwnej

b. Możliwość wyłącznie wczytywania obiektów określonego typu bez możliwości ich zapisu
c. Relację pomiędzy obiektami w kolekcji obiektów wczytywanych z bazy danych narzucają kolejność oraz kierunek ich przeglądania

43. Wybierz stwierdzenia, które odnoszą się do Cassandra Query Language (wiele odp):
    a. Możliwe jest jedynie grupowanie wierszy na poziomie klucza partycji lub na poziomie kolumny grupującej
    b. Wszystkie kolumny klucza głównego muszą być określone w klauzuli WHERE w celu zidentyfikowania konkretnego wiersza, którego to dotyczy
    c. Można łączyć dane operatorem JOIN tylko z dwóch tabel
    d. Nie obsługuje funkcji agregujących - od nowszej wersji weszły (?)
    e. Zapytanie musi odwoływać się tylko do danych umieszczonych w jednej partycji
44. Dopasuj definicję do typu replikacji:
    a. Klienci wysyłają każdy zapis do jednego z kilku węzłów liderów, z których każdy może przyjmować zapisy - Multi-leader replication
    b. Klienci wysyłają każdy zapis do kliku węzłów i odczytują z kilku węzłów równolegle w celu wykrycia i skorygowania węzłów z nieaktualnymi danymi - Leadless replication
    c. Klienci wysyłają wszystkie zapisy do jednego węzła, który wysyła strumień zdarzeń zmiany danych do innych replik - Single-leader replication
    d. Do wyboru: Single-leader replication, Multi-leader replication, Leadless replication
45. Wybierz stwierdzenia, które są prawdziwe (wiele odp):
    a. Bazy NoSQL lepiej się nadają do przechowywania dużych zbiorów danych w środowisku rozproszonym
    b. Nie można wykorzystywać baz relacyjnych w środowisku rozproszonym
    c. Bazy NoSQL są bazami bardziej uniwersalnymi niż bazy relacyjne
    d. Bazy relacyjne są bardziej uniwersalne od baz NoSQL
    e. Model ACID jest wykorzystywany w bazach NoSQL
46. Wybierz stwierdzenia dotyczące baz typu klucz-wartość (key-value) (wiele odp):
    a. Bardzo łatwo jest je skalować
    b. Przykładami baz klucz-wartość są MonogDB i CouchDB
    c. Można w nich przechowywać złożone struktury danych
    d. Są bardzo szybkie
    e. Można w nich przechowywać tylko proste struktury danych
47. Wybierz stwierdzenia, które opisują model danych w Cassandrze (wiele odp):
    a. Wiersze w tabeli mogą składać się z różnych wierszy
    b. Dane całej tabeli są przechowywane w jednym węźle
    c. Dane są przechowywane w wierszach
    d. Dane są przechowywane w kolumnach
    e. Dane tabeli są rozrzucone po różnych węzłach
    f. Wszystkie wiersze tabeli muszą zawierać taki sam zestaw kolumn
48. Wybierz systemy baz danych pracujących w architekturze Leaderless replication (wiele odp):
    a. PostreSQL
    b. Amazon Dynamo - tak (zależy od nazwy dopytać prowadzącego!!!) – amazon dynamoDB!
    c. Cassandra
    d. MongoDB sni
    e. MySQL
49. Wybierz typ(y) partycjonowania wykorzystywane przez Cassandrę (wiele odp):
    a. Sortowanie leksykograficzne (key range partitioning)
    b. Funkcja mieszająca (hash partitioning)

c. W Cassandrze nie jest zaimplementowane partycjonowanie

50. Wybierz stwierdzenia, które odnoszą się do protokołu snitch (wiele odp):
    a. Informacje są wykorzystywane do określania, z których węzłów czytać i do których pisać, a także jak najlepiej dystrybuować repliki, aby zmaksymalizować dostępność w przypadku awarii węzła lub gdy szafa lub centrum danych staje się nieosiągalne
    b. Cassandra decyduje czy węzeł jest włączony czy wyłączony na podstawie tego czy moze sie z nim połączyć poprzez snitch'a co pomaga w optymalnym kierowaniu żądań w ramach klastra
    c. Informują Cassandrę o topologii sieci aby żądania były kierowane efektywnie i pozwalają Cassandrze na dystrybucję replik poprzez grupowanie maszyn w centra danych i szafy
51. Podaj kolejność etapów procesu modelowania danych na potrzeby Cassandry (ułóż w odpowiedniej kolejności, TYLKO 3 ETAPY SĄ WŁAŚCIWE):
    a. Relacyjny model danych
    b. Znormalizowany model danych
    c. Model konceptualny wraz z modelem przypływu danych w aplikacji 1
    d. Logiczny model danych 2
    e. Fizyczny model danych 3
52. Wybierz bazy, które należą do kategorii NoSQL (wiele odp):
    a. Neo4J
    b. Dynamo
    c. BigTable
    d. PostgreSQL
    e. Oracle
    f. Cassandra
    g. MongoDB
    h. MariaDB
53. Wybierz stwierdzenia, które odnoszą się do protokołu gossip (wiele odp):
    a. Co sekundę i wymienia wiadomości o stanie ze wszystkimi innymi węzłami w klastrze.
    b. Informuje Cassandrę o topologii sieci, tak aby żądania były kierowane efektywnie.
    c. Jest protokołem komunikacyjnym typu peer-to-peer, w którym węzły okresowo wymieniają informacje o swoim stanie oraz o innych znanych im węzłach.
    d. Wiadomość ma przypisaną wersję, dzięki czemu podczas wymiany starsze informacje są nadpisywane najbardziej aktualnym stanem dla danego węzła. *(przepisuje z oryginalną składnią Korytkowskiego żeby nie było nieścisłości przy ctrl-f)*
54. Wybierz stwierdzenia, które są prawdziwe w kontekście teorii CAP dotyczącej: (wiele odp)
    a. Teoria Cap dotyczą baz relacyjnych i nierelacyjnych.
    b. Udowodniono, że ta teoria jest już nieaktualna.
    c. Bazy danych mogą spełniać tylko jedno z tych kryteriów.
    d. Wybrane bazy danych mogą spełniać jednocześnie wszystkie trzy kryteria.
    e. Bazy danych mogą spełniać jednocześnie tylko dwa z tych kryteriów.
55. Ułóż Struktury danych wykorzystywane w Cassandrze od największej do najmniejszej: (ułóż od największej do najmniejszej, 6 slotów)
    a. Kolumna
    b. Klaster
    c. Tabela
    d. Wiersz
    e. Przestrzeń kluczy (keyspace)
    f. Partycja
    g. Klaster, Przestrzeń kluczy (keyspace), Tabela, Partycja, Wiersz, Kolumna

56. Wybierz systemy baz danych, które w środowisku rozproszonym pracują w architekturze Single-leader replication: (wiele odp)
    a. Oracle (Oracle Data Guard)
    b. PostgreSQL
    c. Cassandra
    d. MongoDB
    e. MySQL
57. Język SQL trudno stosować do analizy struktur sieciowych ponieważ (jedna odp)
    a. działa zbyt wolno
    b. są one zbyt duże
    c. nie posiada dedykowanych analiz sieciowych
58. Zakładając, że klaster składa się z 11 węzłów umieszczonych w dwóch centrach danych (w Centrum 1 jest 6 węzłów, a w Centrum 2 jest 5 węzłów). W Cassandrze ustawiony jest CONSISTENCY QUORUM, a replication factor ustawiony na 6. Podaj ile węzłów będzie musiało potwierdzić zapis nowych danych do klastra, aby operacja się zakończyła sukcesem. Należy wpisać liczbę:
    a. Q = floor(6/2 +1) = 4
59. Dopasuj definicję typu indeksowania w środowisku rozproszonym (dopasuj, 2 sloty):
    a. Indeksy wtórne przechowywane są w tej samej partycji co klucz główny i wartość. Oznacza to, że tylko jedna partycja musi być aktualizowana przy zapisie, ale odczyt indeksu wtórnego wymaga rozproszenia/ zgromadzenia we wszystkich partycjach. - Indeksy lokalne
    b. Indeksy wtórne są partycjonowane oddzielnie, przy użyciu wartości indeksowych. Wpis w indeksie wtórnym może zawierać rekordy ze wszystkich partycji klucza głównego. Gdy dokument jest zapisywany, kilka partycji indeksu wtórnego musi zostać zaktualizowanych; jednakże odczyt może być obsługiwany w pojedynczej partycji. - Indeksy globalne
    c. Do wyboru: Indeksy globalne, Indeksy lokalne, Indeksy partycjonujące
60. Dobierz typ aplikacji do jej definicji (dopasuj, 6 slotów):
    a. Przechowuje dane tak, aby oni sami lub inna aplikacja mogła je później odnaleźć – Baza danych
    b. Zapamiętują wynik kosztownej operacji, aby przyspieszyć odczyt – cache
    c. Okresowo pobierają i przetwarzają dużą ilość zgromadzonych danych – przetwarzanie wsadowe (batch)
    d. Pozwalają użytkownikom na wyszukiwanie danych według słów kluczowych lub filtrowanie ich na różne sposoby – indeksy
    e. Wysyłają wiadomość do innego procesu, aby została obsłużona asynchronicznie - przetwarzanie strumieniowe (stream)
    f. Do wyboru: bazy danych, przetwarzanie wsadowe (batch), cache, przetwarzanie strumieniowe (stream), indeksy
61. Wybierz struktury danych, które zapisane są w każdej komórce Cassandry (na przecięciu wiersza i kolumny) (wiele odp):
    a. Wartość zgodnie z dale klarowanym typem danych
    b. Typ danych
    c. Czas życia (time to live)
    d. Znacznik czasu (timestamp)
62. Dopasuj definicje: (dopasuj, 2 sloty):
    a. Podniesienie wydajności pojedynczej maszyny - skalowanie wertykalne (pionowe)
    b. Podniesienie wydajności poprzez dodanie większej liczby maszyn – skalowanie horyzontalne (poziome)

c. Do wyboru: skalowanie wertykalne (pionowe), skalowanie horyzontalne (poziome), skalowanie systemowe

63. Wybierz stwierdzenia, które odnoszą się do Cassandry:
    a. Jest to rozproszona baza danych z wieloma liderami
    b. Każdy węzeł może być węzłem koordynującym
    c. Jest to rozproszona baza danych bez lidera
    d. Jest to rozproszona baza danych z jednym liderem
    e. Tylko wybrane węźle (seeds) mogą być węzłami koordynującymi
64. Rozprzestrzenianie informacji w mediach społecznościowych może być modelowane z wykorzystaniem (wiele odp):
    a. modeli epidemiologicznych
    b. modelu SIR
    c. systemów agentowych
65. Podstawą do wyznaczenia miary closeness w analizach sieci są (jedna odp):
    a. najdłuższe ścieżki
    b. najkrótsze ścieżki
    c. połączenia dwukierunkowe
66. W modelu rozprzestrzeniania informacji SIR (jedna odp):
    a. Węzły sieci po wyzdrowieniu nie biorą udziału w procesie propagacji
    b. Węzły sieci po wyzdrowieniu mogą nadal infekować
    c. Węzły sieci po wyzdrowieniu mogą ponownie się zarazić
67. NetLogo to (wiele odp):
    a. Systen modelowania i symulacji wieloagentowych
    b. System grafowych baz danych
    c. Język progamowania oparty na języku Logo
68. Stopień wierzchołka (wiele odp):
    a. jest jedną z miar centralności sieci
    b. jest podstawą do wyznaczania długości ścieżek
    c. określa liczbę sąsiadów węzła
69. Możliwe stany węzłów sieci w modelu rozprzestrzeniania informacji SIR to (jedna odp):
    a. Selected Infected Recovered
    b. Susceptible Interested Recovered
    c. Susceptible Infected Recovered
70. Ekspresyjny język zapytań w grafowych bazach danych, odpowiednik SQLa z baz relacyjnych to (jedna odp):
    a. GraphDB
    b. GraphSQL
    c. Cypher
    d. Cyber
71. Wynikiem zapytań w grafowej bazie danych Neo4j może być (wiele odp):
    a. Wizualizacja grafu
    b. Tabela z numerycznymi i tekstowymi danymi wynikowymi
    c. Ścieżka łącząca zadane węzły (odpowiedzialność grześka)
72. Grafowa baza danych (wiele odp):
    a. zapewnia dedykowane funkcje przetwarzania grafów
    b. zapewnia możliwość wyznaczania najkrótszych ścieżek
    c. wprowadza nowe funkcje do języka SQL
73. W języku Cypher odpowiednikiem polecenia SELECT jest (jedna odp):
    a. Match
    b. Path

c. Search

74. Miara pośrednictwa w strukturach sieciowych określa (jedna odp):
    a. Znaczenie węzła w przesyle informacji pomiędzy segmentami sieci
    b. liczbę kontaktów w transmisji informacji
    c. Liczbę sąsiadów pośredniczących w przesyle informacji

Pytanie **1**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Hadoop został zainspirowany rozwiązaniami wymyślonymi przez

Wybierz jedną odpowiedź:

- a. Cloudera
- b. Google
- c. Amazon
- d. Hortonworks

Pytanie **2**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W kolekcji wynikowej operacji mapToPair

Wybierz jedną odpowiedź:

- a. jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej
- b. jest zawsze więcej elementów niż w kolekcji źródłowej
- c. może być mniej lub więcej elementów niż w kolekcji źródłowej, ale nigdy tyle samo
- d. liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej)
- e. jest zawsze mniej elementów niż w kolekcji źródłowej

Pytanie **3**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Podaj rok pierwszego wydania Hadoop (wersja 0.1.0) - można pomylić się o 2 lata

Odpowiedź: 2006

Pytanie **4**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Ekspresyjny język zapytań w grafowych bazach danych, odpowiednik SQLa z baz relacyjnych to

Wybierz jedną odpowiedź:

- a. GraphSQL
- b. GraphDB
- c. Cypher
- d. Cyber

Pytanie **5**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

NetLogo to

Wybierz wszystkie poprawne:

- [ ] a. System modelowania i symulacji wieloagentowych
- [ ] b. System grafowych baz danych
- [ ] c. Język programowania oparty na języku Logo

Pytanie **6**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W procesie próbkowania Snowball sampling

Wybierz jedną odpowiedź:

- ○ a. Węzły sąsiadujące są odwiedzane wielokrotnie
- ○ b. Węzły sąsiadujące są odwiedzane tylko wtedy, gdy nie zostały odwiedzone w poprzednich iteracjach
- ○ c. Węzły sąsiadujące są odwiedzane tylko wtedy gdy przerwa między iteracjami jest większa niż zadana wartość

Pytanie **7**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście HDFS

Wybierz wszystkie poprawne:

- [ ] a. Klienci zapisują kopie danych bezpośrednio do każdego DataNode.
- [ ] b. Klienci zapisują dane do NameNode.
- [ ] c. W przypadku zatrzymania się węzła NameNode, klaster staje się niedostępny.
- [ ] d. Jest zoptymalizowany do dużych, strumieniowych odczytów plików.
- [ ] e. Nie ma możliwości modyfikacji plików, można tylko dopisać dane na końcu pliku.

Pytanie **8**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście Ranger

Wybierz wszystkie poprawne:

- [ ] a. Wspiera kontrolę dostępu opartą na atrybutach.
- [ ] b. Wykorzystywany w Hadoop od wersji 1.0.
- [ ] c. Wspiera kontrolę dostępu opartą na rolach.
- [ ] d. Umożliwia pracę jednego klastra dla wielu różnych firm (model chmurowy).

Pytanie **9**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Podstawowy komponent Spark to

Wybierz jedną odpowiedź:

- a. Spark RDD
- b. Spark Main
- c. Spark Base
- d. Spark Core

Pytanie **10**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W kolekcji wynikowej operacji flatMap

Wybierz jedną odpowiedź:

- a. liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej)
- b. jest zawsze więcej elementów niż w kolekcji źródłowej
- c. jest zawsze mniej elementów niż w kolekcji źródłowej
- d. jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej
- e. może być mniej lub więcej elementów niż w kolekcji źródłowej, ale nigdy tyle samo

Pytanie **11**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W komponencie Spark SQL

Wybierz jedną odpowiedź:

- a. Import danych jest możliwy z dowolnego systemu baz danych, dla którego istnieje sterownik typu JDBC
- b. Import danych z bazy jest możliwy tylko poprzez plik SQL (dump file)
- c. Import danych jest ograniczony do kilku obsługiwanch przez Spark systemów baz danych

Pytanie **12**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Formaty plików wykorzystywane w HIVE to:

Wybierz wszystkie poprawne:

- [ ] a. ORC
- [ ] b. tekstowy
- [ ] c. Parquet
- [ ] d. Avro
- [ ] e. sekwencyjny

Pytanie **13**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Domyślna liczba utrzymywanych w HDFS kopii danych to (podaj liczbę)

Odpowiedź: 2

Pytanie **14**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W komponencie Spark GraphX podstawową strukturą danych jest

Wybierz jedną odpowiedź:

- a. Skierowany multigraf
- b. Graf skierowany
- c. Graf nieskierowany

Pytanie **15**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Zaznacz właściwe nazwy typów operacji wykonywanych na obiektach RDD w Spark

Wybierz wszystkie poprawne:

- a. Transformacje
- b. Akcje
- c. Transakcje
- d. Funkcje lambda

Pytanie **16**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Próbkowanie sieci oparte na algorytmie Snowball sampling bazuje na

Wybierz jedną odpowiedź:

- a. Włączaniu do próbki tylko jednego sąsiada przetwarzanego węzła
- b. Włączaniu do próbki n sasiadów przetwarzanago węzła
- c. Włączaniu do próbki n losowo wybranych węzłów z sieci

Pytanie **17**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W procesie próbkowania Random Walk

Wybierz wszystkie poprawne:

- a. Możliwe są przeskoki do innych segmentów sieci
- b. Nie są możliwe przeskoki do innych segmentów sieci
- c. Przeskoki do innych segmentów sieci umożliwiają pozyskiwanie próbek z wyizolowanych obszarów do których nie można dotrzeć z innych węzłów

Pytanie **18**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Próbkowanie sieci oparte na algorytmie Random Walk bazuje na

Wybierz jedną odpowiedź:

- a. Włączaniu do próbki n losowo wybranych węzłów z sieci
- b. Włączaniu do próbki n sąsiadów przetwarzanego węzła
- c. Włączaniu do próbki tylko jednego losowo wybranego sąsiada przetwarzanago węzła

Pytanie **19**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście TEZ

Wybierz wszystkie poprawne:

- a. Przyśpiesza obliczenia
- b. Współdziała tylko z HIVE.
- c. Wykorzystuje acykliczny graf skierowany do ustalenia kolejności zadań.
- d. Pośrednie wyniki obliczeń zapisywane są do HDFS.

Pytanie **20**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Hadoop działa w trybie:

Wybierz wszystkie poprawne:

- a. schema-on-read
- b. muszą być zdefiniowane transformacje danych nim dane będą załadowane
- c. schema-on-write
- d. nowe kolumny muszą być zdefiniowane nim dane będą załadowane

Pytanie **21**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Algorytmami szeregowania w YARN są

Wybierz wszystkie poprawne:

- a. FIFO
- b. Capacity
- c. Priority
- d. LIFO
- e. Fair

Pytanie **22**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Dopasuj hasło do opisu w kontekście Zookeepera:

Klient zobaczy ten sam widok systemu, niezależnie od serwera, z którym się łączy.	Atomowość
Aktualizacje od poszczególnych klientów są stosowane w kolejności ich wysyłania.	Spójność sekwencyjna
Serwer z nieświeżymi danymi zostanie zamknięty, zmuszając klienta do przełączenia się na bardziej aktualny serwer.	Aktualność
Aktualizacje przetrwają awarie serwerów.	Trwałość
Aktualizacje albo się udają albo nie.	Pojedynczy obraz systemu

Zadanie 22 – Nie sugerować się odpowiedziami, mają one na celu pokazanie możliwości wyboru.

Pytanie **23**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Język SQL trudno stosować do analizy struktur sieciowych ponieważ

Wybierz jedną odpowiedź:

- a. działa zbyt wolno
- b. są one zbyt duże
- c. nie posiada dedykowanych analiz sieciowych

Pytanie **24**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Dataset API w porównaniu do RDD API zapewnia

Wybierz wszystkie poprawne:

- a. Interfejs programistyczny wykorzystujący język zapytań SQL
- b. Większe możliwości importu danych z różnych źródeł
- c. Większą wydajność obliczeniową

Pytanie **25**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Zaznacz nazwy istniejących komponentów Spark

Wybierz wszystkie poprawne:

- a. Spark Streaming
- b. Spark Graphics
- c. Spark Structural Data
- d. Spark SQL
- e. Spark ALS
- f. Spark GraphX

Pytanie **26**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Apache Spark to (wybierz najbardziej pasującą odpowiedź)

Wybierz jedną odpowiedź:

- a. Biblioteka programistyczna wspomagająca budowę systemów rekomendujących
- b. Biblioteka programistyczna zawierająca metody uczenia maszynowego
- c. Platforma przyspieszająca dostęp do dużych zbiorów danych
- d. Platforma programistyczna do obliczeń rozproszonych

Pytanie **27**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W komponencie Spark Streaming strumien danych przychodzących reprezentowany jest przez typ

Wybierz jedną odpowiedź:

- a. DatasetStream
- b. RDDStream
- c. RDD
- d. DStream

Pytanie **28**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

RDD API zapewnia taką samą wydajność obliczeń bez względu na zastosowany język programowania

Wybierz jedną odpowiedź:

- Prawda
- Fałsz

Pytanie **29**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Zaznacz zadania realizowane przez Zookeepera:

Wybierz wszystkie poprawne:

- a. Wykrywanie awarii jednego z serwerów (robotników lub nadzorcy).
- b. Wybór serwera, który będzie pełnił rolę nadzorcy (mastera).
- c. Wyłączanie węzłów klastra, które nie są chwilowo potrzebne z powodu mniejszego obciążenia zadaniami.
- d. Zarządzanie metadanymi przechowywanymi w znode'ach.
- e. Zarządzanie plikami przechowywanymi w HDFS.

Pytanie **30**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Do metod próbkowania eksploaracyjnego sieci zaliczamy

Wybierz wszystkie poprawne:

- [ ] a. Snow ball sampling
- [ ] b. Random Walk
- [ ] c. Random edge selection

Pytanie **31**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W języku Cypher odpowiednikiem polecenia SELECT jest

Wybierz jedną odpowiedź:

- ( ) a. PATH
- ( ) b. SEARCH
- ( ) c. MATCH

Pytanie **32**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Platforma Apache Kafka wymaga przesyłania komunikatów w formacie:

Wybierz jedną odpowiedź:

- ( ) a. nie wymaga żadnego formatu, może on być dowolny
- ( ) b. XML
- ( ) c. JSON
- ( ) d. AVRO

Pytanie **33**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W komponencie Spark GraphX graf można utworzyć z

Wybierz wszystkie poprawne:

- [ ] a. Samej kolekcji krawędzi
- [ ] b. Kolekcji krawędzi i kolekcji wierzchołków
- [ ] c. Samej kolekcji wierzchołków

Pytanie **34**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Domyślny rozmiar bloku danych w HDFS to

Wybierz jedną odpowiedź:

- a. 256 MB
- b. 256 kB
- c. 8 MB
- d. 32 kB
- e. 32 MB
- f. 128 MB
- g. 128 kB
- h. 8 kB

Pytanie **35**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Apache Spark do prawidłowego działania wymaga dostępu do platformy Hadoop

Wybierz jedną odpowiedź:

- Prawda
- Fałsz

Pytanie **36**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W komponencie Spark SQL zapytania w języku SQL są możliwe

Wybierz jedną odpowiedź:

- a. Dla każdej kolekacji Dataset bez względu na źródło danych
- b. Wyłacznie dla kolekcji Dataset powstałych poprzez zaimportowanie danych z relacyjnej bazy danych

Pytanie **37**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

W komponencie Spark GraphX wartości jakie można przypisać do krawędzi i węzłów grafu to

Wybierz jedną odpowiedź:

- a. Wyłącznie wartości typów prostych (numeryczne, tekstowe, logiczne)
- b. Dowolne wartości (obiekty lub wartości typów prostych)
- c. Wyłącznie wartości liczbowe

Pytanie **38**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Komunikaty przesyłane w Apache Kafka:

Wybierz wszystkie poprawne:

- [ ] a. mogą mieć wiele źródeł (producentów danych)
- [ ] b. mogą mieć wyłącznie jedno źródło (producenta danych)
- [ ] c. mogą mieć wyłącznie jednego odbiorcę (konsumenta danych)
- [ ] d. mogą mieć wielu odbiorców (konsumentów danych)

Pytanie **39**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście Phoenix

Wybierz wszystkie poprawne:

- [ ] a. Obsługuje transakcje.
- [ ] b. Umożliwia analizę danych w trybie OLTP.
- [ ] c. W pełni obsługuję operację JOIN.
- [ ] d. wspiera i wykorzystuje pomocnicze indeksy (secondary indexes).
- [ ] e. Umożliwia dynamiczną rozbudowę schematu kolumn w czasie pracy.

Pytanie **40**

Nie udzielono odpowiedzi

Punkty: 1,00

Oflaguj pytanie

Wynikiem zapytań w grafowej bazie danych Neo4j może być

Wybierz wszystkie poprawne:

- [ ] a. Ścieżka łącząca zadane węzły
- [ ] b. Wizualizacja grafu
- [ ] c. Tabela z numerycznymi i tekstowymi danymi wynikowymi

Apache Spark do prawidłowego działania wymaga dostępu do platformy Hadoop

1/1

Prawda
Fałsz

RDD API zapewnia taką samą wydajność obliczeń bez względu na zastosowany język programowania

1/1

Prawda
Fałsz

Komentarz

Jak jest wykres wydajności dla języków, to python stoi w tyle....

Dopasuj opis do nazwy kategorii:

	Wynik
Protokół bezpiecznego uwierzytelniania: **Kerberos**	1/1
Przechowuje polityki bezpieczeństwa: **Ranger**	1/1
NoSQL'owa baza danych dla Hadoop: **HBase**	1/1
Umożliwia SQL dostęp do danych w Hadoop: **Phoenix**	1/1
Firewall dla hadoop: **Knox**	1/1
Zarządza pracą klastra: **ZooKepper**	1/1
Interfejs graficzny zarządzania klastrem: **Ambari**	1/1

W komponencie Spark GraphX wartości jakie można przypisać do krawędzi i węzłów grafu to

1/1

Wyłącznie wartości liczbowe
Dowolne wartości (obiekty lub wartości typów prostych)

Wyłącznie wartości typów prostych (numeryczne, tekstowe, logiczne)

Przewaga w wydajności Spark względem Hadoop MapReduce wynika głównie z:

1/1

Bardziej efektywnych algorytmów i modeli reprezentacji danych

Zdefiniowania w specyfikacji Spark standardów określających wymagania techniczne klastrów obliczeniowych

Lepszego wykorzystania pamięci RAM maszyn w klastrze obliczeniowym

Zastosowania innego języka programowania

Komentarz

"Jest rozwinięciem idei z Hadoop MapReduce zapewniającym większą wydajność i większe możliwości analityczne"

"Jego bardzo duża szybkość wynika z wykorzystania głównie pamięci RAM w przeciwieństwie do bazującego na zapisach na dysku Hadoop MapReduce"

"Przewaga w wydajności wynika także z zastosowanych metod reprezentacji danych w silniku obliczeniowym bazujących na skierowanych grafach acyklicznych (DAG Engine)"

W modelu rozprzestrzeniania SIR

1/1

Węzły sieci po wyzdrowieniu mogą ponownie się zarazić
Węzły sieci po wyzdrowieniu mogą nadal infekować
Węzły sieci po wyzdrowieniu nie biorą udziału w procesie propagacji

Komentarz

SIR = Susceptible, Infectious, or Recovere

Ekspresyjny język zapytań w grafowych bazach danych, odpowiednik SQLa z baz relacyjnych to

1/1

Cyber
GraphSQL
GraphDB
Cypher

Próbkowanie sieci oparte na algorytmie Snowball sampling bazuje na

1/1

Włączaniu do próbki n losowo wybranych węzłów z sieci
Włączaniu do próbki n sąsiadów przetwarzanego węzła

Włączaniu do próbki tylko jednego sąsiada przetwarzanego węzła

Próbkowanie sieci oparte na algorytmie Random Walk bazuje na:

1/1

Włączaniu do próbki n losowo wybranych węzłów z sieci
Włączaniu do próbki n sąsiadów przetwarzanego węzła
Włączaniu do próbki tylko jednego losowo wybranego sąsiada przetwarzanego węzła

W kolekcji wynikowej operacji mapToPair

1/1

jest zawsze mniej elementów niż w kolekcji źródłowej
może być mniej lub więcej elementów niż w kolekcji źródłowej, ale nigdy tyle samo
liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej)
jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej

jest zawsze więcej elementów niż w kolekcji źródłowej

Komentarz

mapToPair(func)
Mapuje elementy ze źródłowej RDD na elementy typu klucz-wartość. Każdy
z elementów powstałej "dwójki" może mieć swój odrębny typ.

Moim zdaniem mniej ale nie jestem pewien

W komponencie Spark Streaming strumień danych przychodzących reprezentowany jest przez typ

1/1

DatasetStream
RDDStream
RDD
DStream

W komponencie Spark SQL

1/1

Import danych z bazy jest możliwy tylko poprzez plik SQL (dump file)
Import danych jest ograniczony do kilku obsługiwanych przez Spark systemów baz danych
Import danych jest możliwy z dowolnego systemu baz danych, dla którego istnieje sterownik typu JDBC

Komentarz

-> Import danych z systemu zarządzania bazą danych
z wykorzystaniem interfejsu JDBC
->Wymagane jest dołączenie do projektu
sterownika JDBC dla używanej bazy danych

Podaj rok pierwszego wydania Hadoop (wersja 0.1.0) - można pomylić się o 2 lata

1/1

2006

Komentarz

April 2006
In March 2006, Owen O'Malley was the first committer to add to the Hadoop project; Hadoop 0.1.0 was released in April 2006. It continues to evolve through contributions that are being made to the project.

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście HIVE?

1/1

Dane tabel są przechowywane w katalogu w HDFS

Jest w pełni zgodny z ACID

Dane są przechowywane w Metastore

Wykorzystywany jest mechanizm optymalizacji zapytań

Komentarz

Optymalizacja -> internet + grafy z prezentacji (istnieje optymalizator w sterowniku)

Tabela jest przechowywana w katalogu w HDFS.

Meta dane są przechowywane w metastore (nwm czy dane) CHEATS ;-;

Wskaż cechy hadoop

1/1

ma 100% zgodność z SQL
działa w trybie OLAP
działa w trybie ACID

ma elastyczną strukturę danych

umożliwia złożone przetwarzanie danych

jest skalowalny

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście YARN?

1/1

Pozwala na pracę wielu silników przetwarzających dane na jednym klastrze

Zarządca zasobów działa w węzłach podrzędnych (DataNode)

Zadania odpalane są w kontenerach

Jest jeden Application Master dla wszystkich zapytań

Jest warstwą przetwarzania Hadoopa

Które języki programowania posiadają oficjalne wsparcie dla Apache Spark

1/1

Perl
Swift
Java
Kotlin
Scala
Python

Domyślny rozmiar bloku danych w HDFS to

1/1

256 kB
32 kB
8 MB
8 kB
256 MB
128 MB
32 MB
128 kB

Sqoop umożliwia:

1/1

Transfer danych do HBase
Transfer danych między Hadoop, a wieloma bazami relacyjnymi
Transfer danych tylko między Hadoop i mysql

Podstawowy komponent Spark to

1/1

Spark Base
Spark Main
Spark Core
Spark RDD

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście Ranger

1/1

Umożliwia pracę jednego klastra dla wielu różnych firm (model chmurowy)
Wspiera kontrolę dostępu opartą na rolach

Wykorzystywany w Hadoop od wersji 1.0

Wspiera kontrolę dostępu opartą na atrybutach

Komentarz

Enhanced support for different authorization methods - Role based access control, attribute based access control etc.

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście TEZ?

1/1

Przyśpiesza obliczenia

Współdziała tylko z HIVE

Pośrednie wyniki obliczeń zapisywane są do HDFS

Wykorzystuje acykliczny graf skierowany do ustalenia kolejności zadań

Komentarz

Przyspiesza obliczenia -> "Silnik wykonujący lepiej operacje MapReduce (ok. 10x szybciej)"
"DAG - (Directed- Acyclic-Graph) – skierowany graf acykliczny"
Reszty nie wiem

W komponencie Spark SQL zapytania w języku SQL są możliwe

1/1

Dla każdej kolekcji Dataset bez względu na źródło danych

Wyłącznie dla kolekcji Dataset powstałych przez zaimportowanie danych z relacyjnej bazy danych

Dataset API zapewnia taką samą wydajność obliczeń bez względu na zastosowany język programowania

1/1

Prawda

Fałsz

Komentarz

W wykladzie spark SQL gdzie jest dataset:
"Przetwarzanie jest niezależne od użytego API czy języka programowania"

W kolekcji wynikowej operacji flatMap

1/1

jest zawsze mniej elementów niż w kolekcji źródłowej
może być mniej lub więcej elementów niż w kolekcji źródłowej, ale nigdy tyle samo
jest zawsze dokładnie tyle samo elementów co w kolekcji źródłowej
liczba elementów może być dowolna (mniejsza lub większa lub taka sama jak w kolekcji źródłowej)

jest zawsze więcej elementów niż w kolekcji źródłowej

Apache Spark to (wybierz najbardziej pasującą odpowiedź)

1/1

Biblioteka programistyczna wspomagająca budowę systemów rekomendujących
Platforma przyspieszająca dostęp do dużych zbiorów danych
Platforma programistyczna do obliczeń rozproszonych

Biblioteka programistyczna zawierająca metody uczenia maszynowego

Które założenia bezpieczeństwa w Hadoop są prawdziwe?

1/1

Komunikacja HDFS i MapReduce nie będzie działać w niezaufanych sieciach

Hadoop tworzy konta dla użytkowników

Domyślnie zadanie w klastrze może trwać maksymalnie 7 dni

Użytkownicy muszą mieć dostęp do kont root w klastrze

Dostęp do HDFS będzie autoryzowany za pomocą tokenów

Komentarz

Wykład -> "Komunikacja HDFS i MapReduce nie będzie działać w niezaufanych sieciach"
-> "Zadanie Hadoop będzie działać nie dłużej niż 7 dni (konfigurowalne) w klastrze MapReduce lub dostęp do HDFS z zadania zakończy się niepowodzeniem"
-> Bilety Kerberos nie będą przechowywane w zadaniach MapReduce i nie będą dostępne dla zadań zadania. Dostęp do HDFS będzie autoryzowany za pomocą tokenów

NetLogo to:

1/1

System grafowych baz danych

System modelowania i symulacji wieloagentowych

Język programowania oparty na języku Logo

Komentarz

Wiki - NetLogo to język programowania i zintegrowane środowisko modelowania matematycznego.
Eng wiki - NetLogo is a programming language and integrated development environment (IDE) for agent-based modeling.

Które z poniższych stwierdzeń są PRAWDZIWE w kontekście MapReduce?

1/1

Wartości związane z danym kluczem zawsze trafią do tego samego reduktora

Najpierw odbywa się faza redukcji a następnie mapowania

Maper odczytuje dane w postaci par klucz/wartość

Hadoop stara się zapewnić, że mapery działają na węzłach, które przechowują lokalnie swoją część danych

Wynik pracy węzła jest na bieżąco wysyłany do kolejnego węzła

Zaznacz właściwe nazwy typów operacji wykonywanych na obiektach RDD w Spark

1/1

Akcje

Transakcje

Transformacje

Funkcje lambda

Zagadnienia egzaminacyjne

1. Czym są duże zbiory danych: charakterystyka, obszary zastosowań.
2. Hadoop: charakterystyka, zastosowania, elementy składowe, zapewnienie niezawodności działania.
3. BigData vs. relacyjne bazy danych.
4. HDFS: charakterystyka, redundancja danych w HDFS, szybkość zapisu i odczytu, niezawodność, zadania Namenode i Datanode.
5. YARN: charakterystyka, tryby pracy.
6. Paradygmat MapReduce: charakterystyka, zasada działania, możliwości i ograniczenia.
7. HIVE: charakterystyka, zastosowania, możliwości i ograniczenia HiveQL, Sqoop.
8. TEZ: charakterystyka, mechanizmy które pozwoliły na przyśpieszenie pracy Hive: ORC, LLAP, wektoryzacja, DAG – skierowane grafy acykliczne.
9. Hbase: charakterystyka, zasada działania, możliwości i ograniczenia, zadania Master i Regionserver. (tylko studia stacjonarne)
10. Phoenix: charakterystyka, zasada działania. (tylko studia stacjonarne)
11. Bezpieczeństwo Hadoop: architektura, identyfikacja, uwierzytelnianie, autoryzacja, audytowanie. Kerberos, Apache Ranger, Apache Knox. (tylko studia stacjonarne)
12. Apache Spark: charakterystyka, zastosowanie, zasada działania, podstawowe moduły, obsługiwane języki programowania, różnice względem Hadoop MapReduce, RDD (transformacje, akcje), nowe API oparte o DataSet i DataFrame, przykłady zastosowania Spark w obszarze uczenia maszynowego (Machine Learning).
13. Sposoby oceny jakości próbkowania sieci.
14. Różnice między agregacją danych z sieci a próbkowaniem.
15. Główne strategie próbkowania sieci homogenicznych.
16. Rożnice między Random Walk a Snowball Sampling.
17. Główne cechy języka Cypher.
18. Omówić korzyści wynikające z zastosowania grafowych baz danych do przeszukiwania połączeń lotniczych w porównaniu do relacyjnych baz danych.

1. Czym są duże zbiory danych: charakterystyka, obszary zastosowań.

Big Data to termin opisujący dużą ilość danych - zarówno ustrukturyzowanych, jak i nieustrukturyzowanych, których analiza nie jest trywialna, lecz może prowadzić do lepszych decyzji i strategicznych ruchów biznesowych.

Charakterystyka:
• **Ilość (volume) -** dane zbierane z różnorodnych źródeł: transakcje biznesowe, media społecznościowe, dane z sensorów, dane wymieniane między urządzeniami.

• **Szybkość (velocity)** - dane powstają i są dostarczane niezwykle szybko i muszą być obsługiwane w odpowiednim czasie, często zbliżonym do trybu czasu rzeczywistego.
• **Różnorodność (variety)** - dane przychodzą w różnych formatach – od ustrukturyzowanych, numerycznych danych w tradycyjnych bazach danych do niestrukturalnych dokumentów tekstowych, email, video, audio, danych znaczników magazynowych lub transakcji finansowych.

• **Zmienność (variability)** - przepływy danych mogą podlegać dużym wahaniom okresowym. Czy w mediach społecznościowych jakiś temat jest szczególnie popularny? Czasami trudno jest zarządzać danymi napływającymi w trakcie szczytu dziennego, sezonowego czy wywołanego konkretnym wydarzeniem. Jest to jeszcze trudniejsze w przypadku danych niestrukturalnych.

• **Złożoność (complexity)** - dane napływają do nas z wielu różnych źródeł. Łączenie, dopasowywanie, oczyszczanie i przekształcanie danych w różnych systemach to zadania wymagające dużego wysiłku. Niemniej jednak łączenie i zestawianie relacji, hierarchii i różnorodnych powiązań między danymi jest niezbędne.

Obszary zastosowań:

- **Banki**, dla których kluczowe jest zrozumienie klienta i podniesienie jego satysfakcji z oferowanych mu usług, ale z drugiej strony konieczne jest minimalizowanie ryzyka i redukcja potencjalnych nadużyć
- **Produkcja,** podniesienie jakości produktów, zwiększenie wydajności produkcji oraz ograniczenie strat
- **Handel detaliczny,** analiza danych pozwala na budowanie trwałych relacji z klientami, uzyskanie optymalnych sposobów dotarcia do klientów, odkrycie sposobów na odzyskanie utraconych szans sprzedażowych
- **Ochrona zdrowia,** efektywne zarządzanie big data pozwala służbie zdrowia odkryć nieznane zależności oraz polepszyć obsługę pacjenta.
- **Sektor publiczny** - usprawnienie zarządzania, optymalizacja kosztów, zwiększenie jakości obsługi obywateli oraz przeciwdziałanie przestępczości.
- **Edukacja -** dzięki analizie big data można identyfikować zagrożenia dla uczniów, pomagać studentom w wyborze właściwej ścieżki edukacji oraz usprawnić system oceny i wsparcia nauczycieli.

2. Hadoop: charakterystyka, zastosowania, elementy składowe, zapewnienie niezawodności działania.

Hadoop to narzędzie stworzone przez fundację Apache do zarządzania gigantycznymi zasobami danych i przetwarzania ich. Jest to oprogramowanie open source.

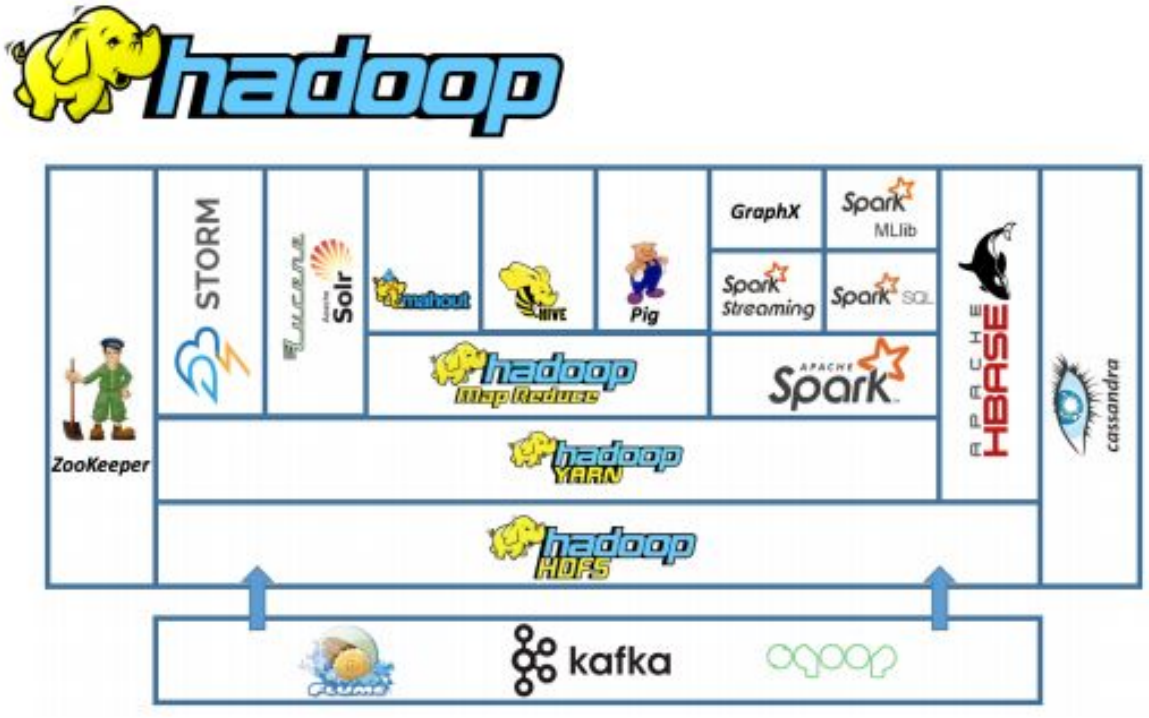
Charakterystyka:

- posiada niezawodny magazyn HDFS i system analizy MapReduce

- wysoka skalowalność (liniowa), dzięki temu może zawierać nawet tysiące serwerów, które można stale do niego dodawać
- wysoka opłacalność, może pracować nawet z tanim sprzętem i nie wymaga sprzętu high-end
- jest elastyczny i może przetwarzać dane ustrukturyzowane, jak i nieustrukturyzowane
- wbudowana odporność na uszkodzenia i replikacja na wielu węzłach (automatyczna), nawet w przypadku awarii węzła można dane odczytać z innego
- działa na zasadzie jednokrotnego zapisu i wielokrotnego odczytu
- jest zoptymalizowany dla dużych i ogromnych danych i nie nadaje się do małych zestawów danych
- zrównoleglanie operacji przetwarzania danych
- uniezależnienie od producenta sprzętu
- brak ograniczeń rozmiarowych plików
- umożliwia pracę na SQL + inny język programowania
- przechowywanie danych w dowolnej postaci
- archiwizacja danych do postaci oszczędzania miejsca wraz z możliwościa ich bezproblemowego odtworzenia w dowolnym momencie

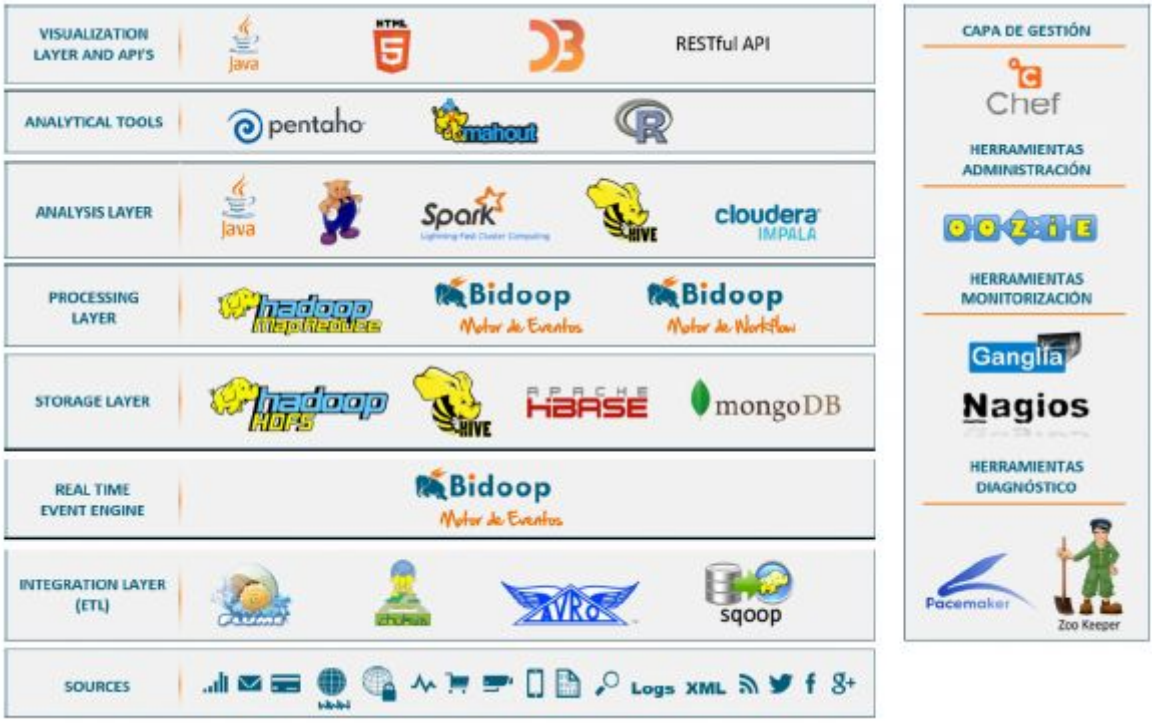
Zastosowania:
- Analityka
- Przechowywanie danych
- Przetwarzanie danych (np pliki dziennika, dzienne raporty itp)
- Analiza treści tekstowych, graficznych, audio i wideo
- systemy rekomendacji

Elementy Składowe:



Dodatkowo:

- **Hadoop Common –** biblioteki i narzędzia używane przez pozostałe moduły
- **Hadoop Distributed File System (HDFS) –** rozproszony system plików
- **YARN –** platforma do zarządzania zasobami klastra
- **MapReduce –** implementacja paradygmatu MapReduce do przetwarzania dużych ilości danych
- Namenode
- JobTracker
- Task Tracer
- Datanode
- Resource Manager
- App Manager
- Container



Warstwy (od dołu):

- Źródła
  - logi, XML, poczta, media społecznościowe
- integracji (ETL)
  - sqoop, flume, chukwa, avro
- Silnik zdarzeń czasu rzeczywistego
  - Bidoop (Motor de eventos)
- Przechowywania danych
  - HDFS, HIVE, HBASE, mongoDB
- Przetwarzania
  - MapReduce, Bidoop (Motor de eventos + (Motor de Workflow)
- Analizy
  - Java, Apache Pig, Spark, HIVE, Apache Impala
- Narzędzia analityczne

    - Pentaho, Apache Mahout
- Wizualizacji oraz API
    - Java, HTML, D3 (biblioteka JS), RESTful API

Zapewnienie niezawodności działania:

Zookeeper i główny NameNode (master) stale czuwają nad bezpieczeństwem danych i działaniem całego zestawu klastrów i hadoopa. W momencie kiedy którychś z serwerów z danymi padnie czy to podczas działań na danych czy podczas rozdzielania danych przez NameNode, Zookeeper natychmiast przełącza na inny dostępny, działający serwer zawierający te same dane.

Hadoop zawsze replikuje zapisane konkretne dane w 2 dodatkowych kopiach na inne DataNode, w przypadku awarii jednego z nich, sprawdza jakie dane się tam znajdowały i replikuje te dane na inne dostępne DataNode z wolnymi zasobami.

3. BigData vs. relacyjne bazy danych.

Relacyjna BD:

- OLAP (Online analytical processing) - operowanie na wielowymiarowych bazach danych
- Transakcje - ACID (niepodzielność, spójność, izolacja, trwałość) - zapewnia że transakcje albo dojdą do skutku ze 100% wykonaniem, albo w ogóle nie zostaną wykonane w przypadku błędu.
- 100% SQL
- Schema-on-write:
    - Struktura danych musi być zdefiniowana przed zapisem
    - Muszą być zdefiniowane transformacje danych do wewnętrznej struktury bazy
    - Nowe kolumny muszą być zdefiniowane (dodane) nim dane będą załadowane do bazy
    - Szybki odczyt
    - Standaryzacja

Hadoop :

- Elastyczna struktura danych
- Skalowalność
- Złożone przetwarzanie danych
- Schema-on-read
    - Dane są kopiowane do pliku bez żadnej transformacji.
    - Mechanizm serializacji jest wykorzystany do odczytu danych aby uzyskać dostęp do kolumn
    - Nowy typ danych może przyjść w dowolnym momencie i wykorzystany po aktualizacji mechanizmów serializacji

- szybki zapis
- elastyczność

4. HDFS: charakterystyka, redundancja danych w HDFS, szybkość zapisu i odczytu, niezawodność, zadania Namenode i Datanode.

charakterystyka

- rozproszony system plików stworzony dla Hadoop (ale nie wymagany przez Hadoop)
- możliwość pracy na dowolnym sprzęcie (w sensie wydajnościowym/architektonicznym)
- przeznaczenie do pracy z ogromnymi rozmiarami plików
- nie ma możliwości zmiany w plikach, tylko zapis/usunięcie/odczyt
- duży rozmiar bloku (64MB/128MB?)

redundancja danych w HDFS

- dzielenie plików w bloki, które są replikowane na różnych klastrach
- domyślna redundancja pliku to 3 razy.

szybkość zapisu i odczytu

? przypuszczam, że odczyt jest odpowiednio szybszy coś jak raid 1, zapis bez zmian, przy dużych plikach, małe pliki idą wolno przez rozmiar bloku ?

niezawodność

- wysoka odporność na uszkodzenia poprzez dzielenie plików w bloki, które są replikowane na różnych klastrach

zadania Namenode i Datanode

NameNode, znany jako Master node, Zarządza przestrzenią nazw systemu plików i wykonuje operacje takie jak otwieranie, zamykanie i zmiana nazw plików i katalogów. Utrzymuje drzewo systemu plików i metadane (liczbę bloków danych *, replik itp.) Dla wszystkich plików i katalogów w drzewie.

DataNode, Zwykle działa na osobnej maszynie fizycznej. Jego rolą jest okresowe łączenie FsImage i EditLogs z NameNode. Zapobiega to zbyt dużemu dziennikowi edycji. Przechowuje również kopię scalonego FsImage w trwałym magazynie, którego można użyć w przypadku awarii NameNode.
FsImage: Trwały punkt kontrolny metadanych systemu plików.
EditLogs: Zawiera wszystkie ostatnie modyfikacje wprowadzone w systemie plików w odniesieniu do najnowszego FsImage.

5. YARN: charakterystyka, tryby pracy.

decentralizuje wykonywanie i monitorowanie zadań przetwarzania. Odpowiedzialne za to komponenty:

ResourceManager: który przyjmuje zgłoszenia zadań od użytkowników, planuje zadania i przydziela im zasoby;

NodeManager: który jest zainstalowany w każdym węźle i działa jako agent monitorowania i raportowania ResourceManagera

ApplicationMaster: który jest tworzony dla każdej aplikacji w celu negocjacji zasobów i współpracy z NodeManagerem w celu wykonywania i monitorowania zadań

6. Paradygmat MapReduce: charakterystyka, zasada działania, możliwości i ograniczenia.

MapReduce to platforma do przetwarzania równoległego dużych zbiorów danych w klastrach komputerów stworzona przez firmę Google. (wiki)

Zasada działania

(wiki) Operacje realizowane są podczas dwóch kroków:

Krok "map" - węzeł nadzorczy (master node) pobiera dane z wejścia i dzieli je na mniejsze podproblemy, po czym przesyła je do węzłów roboczych (worker nodes). Każdy z węzłów roboczych może albo dokonać kolejnego podziału na podproblemy, albo przetworzyć problem i zwrócić odpowiedź do głównego programu.

Krok "reduce" - główny program bierze odpowiedzi na wszystkie podproblemy i łączy je w jeden wynik - odpowiedź na główny problem.

Możliwości i ograniczenia

(wiki)Główną zaletą MapReduce jest umożliwienie łatwego rozproszenia operacji. Zakładając, że każda z operacji "map" jest niezależna od pozostałych, może być ona realizowana na osobnym serwerze.

MapReduce mapuje i redukuje sekwencje operacji na rozproszonym zestawie serwerów.

Map: mapuje swoje dane do par klucz-wartość, przekształca je i filtruje. Następnie przypisuje dane do węzłów do przetwarzania.

Reduce: agreguje te dane do mniejszych zbiorów danych. Dane z kroku zmniejszania są przekształcane do standardowego formatu klucz-wartość - w którym klucz działa jako identyfikator rekordu.
definicja z
https://www.dummies.com/programming/big-data/data-science/the-mapreduce-programming-paradigm/

7. HIVE: charakterystyka, zastosowania, możliwości i ograniczenia HiveQL, Sqoop.

Charakterystyka

Apache Hive to system hurtowni danych dla Apache Hadoop. Hive umożliwia tworzenie podsumowań danych, tworzenie zapytań i analizę danych. Zapytania Hive są pisane w HiveQL, który jest językiem zapytań podobnym do SQL. Hive pozwala wyświetlać strukturę danych o dużych rozmiarach. /* Po zdefiniowaniu struktury możesz użyć HiveQL do zapytania na danych bez znajomości Java lub MapReduce. ?*/

Zastosowania

Hive powinien być używany do analitycznych zapytań w danych gromadzonych przez pewien okres czasu. np. oblicz trendy, podsumuj logi witryny. Nie można go używać do zapytań w czasie rzeczywistym (real time queries).

Możliwości i ograniczenia HiveQL:

Możliwości:

1. Hive umożliwia wykonywanie zapytań na dużych zbiorach nieustrukturyzowanych danych, które nie muszą być wcześniej odpowiednio sformatowane. Dane mogą mieć różnorodne formaty, od nieustrukturyzowanych plików, po pliki JSON i tabele HBase.
2. Zawiera takie instrukcje SQL jak m.in. SELECT, INSERT, GROUP BY, ORDER BY, SORT BY, JOIN, UNION oraz obsługuje wiele typów danych, m.in. INT, BOOLEAN, FLOAT, DOUBLE, STRING, BINARY, DECIMAL.
3. Pozwala na wyświetlanie analiz dla danych np. w postaci wykresów.

Ograniczenia:

HiveQL nie wspiera:

1. instrukcji związanych z transakcjami ACID (INSERT, UPDATE, DELETE),
2. podzapytań SQL w klauzulach WHERE, HAVING,
3. rozszerzonej składni JOIN,
4. instrukcji dotyczących bezpieczeństwa (GRANT, itp.)
5. typu danych CHAR

Sqoop

Narzędzie działające w wierszu poleceń przeznaczone do przesyłania danych między klastrami Hadoop a relacyjnymi bazami danych (np. MySQL i Oracle)

8. TEZ: charakterystyka, mechanizmy które pozwoliły na przyspieszenie pracy Hive: ORC, LLAP, wektoryzacja, DAG – skierowane grafy acykliczne.

TEZ: Charakterystyka:

Apache Tez to framework, który wykorzystywany jest jako silnik optymalizacji wywoływania zadań MapReduce dla skryptów Pig i Hive. Dzięki wykorzystaniu Wykonanie skryptów Hive z użyciem Tex może być nawet kilkanaście razy szybsze.

Mechanizmy, które pozwoliły na przyspieszenie pracy Hive:

ORC - Optimized Row Column

1. kolumnowy format danych dla skomplikowanych typów danych
2. Od Hive 0.11 - przechowywanie danych
3. Wsparcie dla Pig i MapReduce przez HCat
4. Dwa poziomy kompresji: generyczny i lightweight type-specific (nie wiem jak to przetłumaczyć, żeby ładnie brzmiało - lekki i zależny od typu? xd

LLAP - Live Long and Process

1. Zestaw trwałych (długo-żyjących) procesów (demonów) do buforowania danych, zarządzania optymalizacją JIT i eliminacji większości kosztów uruchamiania.
2. Małe zapytania są przetwarzane głównie przez demony bezpośrednio, natomiast każde duże żądanie zostanie wykonane za pośrednictwem Hive
3. Unikanie długich czasów odpowiedzi - poprawia wydajność

Wektoryzacja

Metoda optymalizacji zapytań Hive, dzięki której Hive może przetwarzać partie 1024 wierszy na raz, zamiast przetwarzać każdy wiersz osobno, co znacznie przyspiesza wykonywanie prostych operacji.

DAG – skierowane grafy acykliczne.

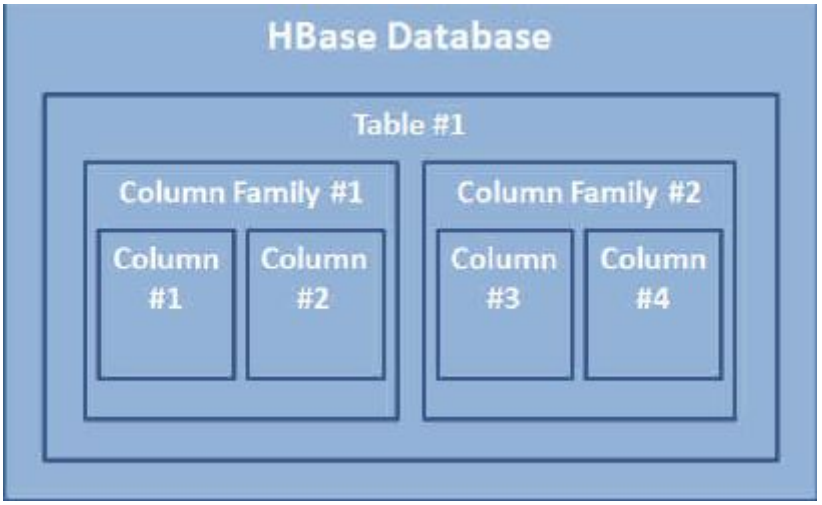
DAG to zbiór wierzchołków, w którym każdy wierzchołek wykonuje fragment zapytania lub skryptu. Bezpośrednie połączenia między wierzchołkami określają kolejność ich wykonywania.

9. Hbase: charakterystyka, zasada działania, możliwości i ograniczenia, zadania Master i Regionserver

- Hbase jest rozproszoną bazą danych zbudowaną na bazie HDFS.
- Bazuje na rozwiązaniu Google Bigtable
- Hbase jest aplikacją Hadoop przeznaczoną do losowego dostępu do bardzo dużych zbiorów danych w czasie rzeczywistym.
- Skaluje się ona liniowo poprzez dodawanie węzłów
- **Hbase nie jest relacyjna i nie obsługuje SQL!**

- Jest w stanie zrobić to, czego RDBMS nie jest w stanie zrobić(obsługiwać bardzo duże słabo zapełnione tabele na klastrach wykonanych z typowego sprzętu)

- Dane są przechowywane w tabelach
- Tabele składają się z wierszy i kolumn
- Zawartość komórki jest nie interpretowaną tablicą bajtów
- Kolumny wierszy są pogrupowane w rodziny kolumn
- Wszyscy jej członkowie mają wspólny prefiks
- Kolumna rodzina i kwalifikator są zawsze oddzielone znakiem dwukropka (:)
- Rodziny kolumn tabeli muszą być określone z góry jako część definicji schematu tabeli, ale nowi członkowie rodzin kolumn mogą być dodawani na żądanie



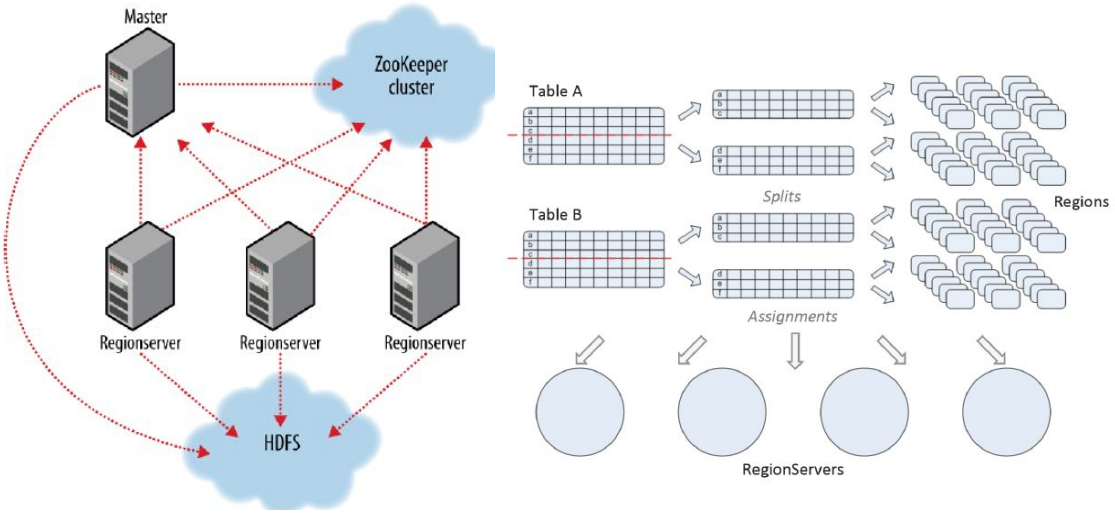
- Komórki tabeli - przecięcie wiersza i kolumny - są wersjonowane. Domyślnie parametrem wesjonowania jest znacznik czasu automatycznie przypisany przez Hbase w momencie wstawiania komórki.
- Fizycznie, wszyscy członkowie rodziny kolumn są przechowywani razem w systemie plików(hfile)
- **HBase jest więc bazą rodzin kolumn!**
- Optymalizacja bazy przechowywania są wykonywane na poziomie rodzin kolumn, zaleca się aby wszyscy członkowie rodziny kolumn mieli ten sam ogólny wzór dostępu i tę samą charakterystykę rozmiaru. W przypadku tabeli zdjęć dane zdjęciowe, które są duże, są przechowywane w osobnej rodzinie kolumn od rodziny metadane, które są znacznie mniejsze.
- Wiersze w tabeli zawsze są posortowane wg klucza
- Aktualizacje wierszy są atomowe, bez względu na to, ile kolumn wierszy składa się na transakcję na poziomie wiersza. Dzięki temu model blokady jest prostu.
- Tabele są automatycznie dzielone poziomo przez Hbase na regiony.
- Każdy region składa się z podzbioru wierszy tabeli
- Region jest oznaczony przez tabelę, do której należy jej pierwszy wiersz(włącznie), i ostatni wiersz(wyłącznie)
- Początkowo tabela składa się z jednego regionu, ale w miarę jak region rośnie i przekracza konfigurowalny próg wielkości, rozbijana jest na 2 nowe regiony o mniej więcej równej wielkości. Dopóki ten pierwszy podział nie zostanie dokonany całe

obciążenie będzie skierowane na pojedyńczy serwerów hostującemu oryginalny region.

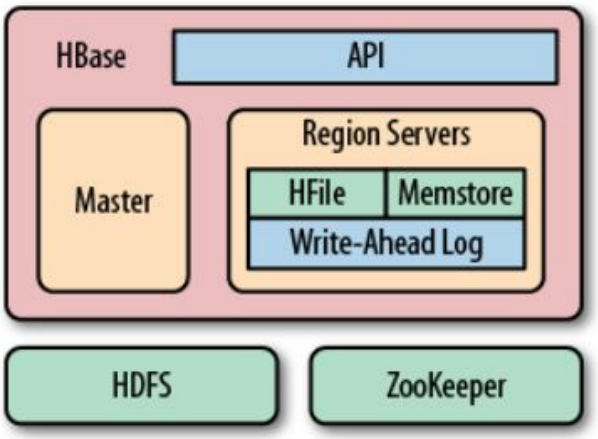
- Czyli w miarę jak tabela rośnie, liczba jej regionów rośnie

Zadania Master i Regionserver:

- Regiony to jednostki, które są rozmieszczone na klastrze HBase
- W ten sposób, tabela, która jest zbyt duża dla jednego serwera może być przenoszona przez klaster serwerów, z każdym węzłem hostującym podzbiór wszystkich regionów tabeli
- Jest to również sposób, za pomocą którego obciążenie tabeli jest dystrybuowane
- Zestaw posortowanych regionów składa się na całą zawartość tabeli



- Serwery regionalne mogą być dodawane lub usuwane podczas pracy systemu, aby dostosować się do zmieniającego się obciążenia pracą.
- Master jest odpowiedzialny za przydzielanie regionów do serwerów regionalnych i używa Apache ZooKeeper, niezawodnej wysoce dostępnej, stałej i rozproszonej usługi koordynacji, aby ułatwić to zadanie.
- Serwer master jest również odpowiedzialny za obsługę równoważenia obciążenia regionów na serwerach regionalnych, aby rozładować obciążone serwery i przenieść regiony do mniej obciążonych
- Serwer master nie jest częścią faktycznej ścieżki przechowywania lub pobierania danych. Negocjuje on równoważenie obciążenia i utrzymuje stan klastra, ale nigdy nie świadczy żadnych usług związanych z danymi ani dla serwerów regionalnych, ani dla klientów, dlatego w praktyce jet lekko obciążony. Ponadto zajmuje się zmianami schematów i innymi operacjami na metadanych, takimi jak tworzenie tabel i rodzin kolumn.
- Serwery regionalne są odpowiedzialne za wszystkie żądania odczytu i zapisu dla wszystkich regionów, które obsługują, a także dzielą regiony, które przekroczyły skonfigurowane progi wielkości regionów. Klienci komunikują się z nimi bezpośrednio w celu obsługi wszystkich operacji związanych z danymi.



10. Phoenix: charakterystyka, zasada działania

Apache Phoenix umożliwia analizę OLTP i analizę operacyjną w Hadoopie dla aplikacji o niskich opóźnieniach poprzez połączenie zalet obu światów:

- moc standardowych interfejsów API SQL i JDBC z pełnymi możliwościami transakcji ACID
- schema-on-read ze świata NoSQL poprzez wykorzystanie HBase jako nośnika danych

Apache Phoenix jest w pełni zintegrowany z innymi produktami Hadoop, takimi jak Spark, Hive, Pig, Flume i Map Reduce.

Dzisiejszym wyzwaniem jest połączenie skali i wydajności NoSQL z łatwością obsługi SQL

Phoenix daje ponad standardowe możliwości SQL:

- Widoki(views) tylko do odczytu na istniejących danych z HBase
- Dynamiczną rozbudowę schematu kolumn w czasie pracy
- Schemat bazy jest wersjonowany - dzięki właściwością HBase - możliwość zapytań retrospektywnych z wykorzystaniem wcześniejszych wersji metadanych

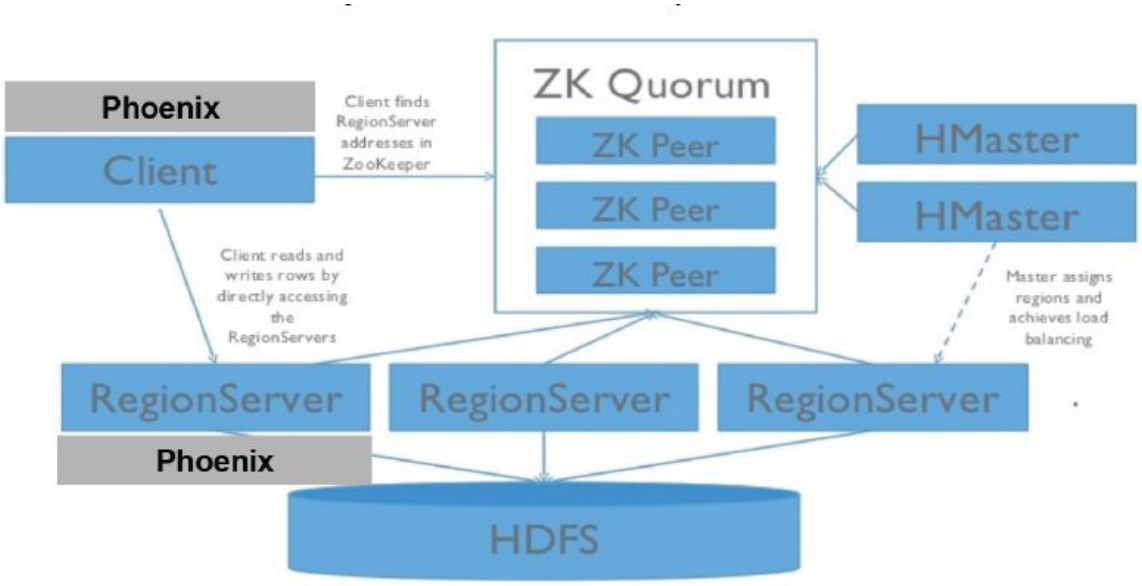
Działają: CREATE, DROP, ALTER TABLE, INSERT, UPADATE, DELETE, SELECT, WHERE, HAVING, GROUP BY, ORDER BY, LIMIT, VIEW, INDEX
W ograniczonym zakresie działa JOIN
Nie działa TRANSACTION

Dostęp do danych przez Phoenix może być o wiele szybszy niż bezpośrednio przez HBase ponieważ:

- Phoenix zrównolegla zapytania na podstawie zgromadzonych statystyk(HBase nie wie jak podzielić zapytania, musi przeskanować cały region)
- Phoenix wspiera i wykorzystuje pomocnicze indeksy(secondary indexes)
- Phoenix wykorzystuje różnego typu optymalizacje, np. wykorzystanie koprocesorów

Biblioteka jak Phoenix'a zainstalowana jest na RegionServer Hbase

Sterownik JDBC - Phoenix jest zainstalowany na kliencie



11. Bezpieczeństwo Hadoop: architektura, identyfikacja, uwierzytelnianie, autoryzacja, audytowanie. Kerberos, Apache Ranger, Apache Knox.

- Wczesne wersje Hadoopa zakładały wykorzystanie klastrów HDFS i MapReduce przez grupę współpracujących użytkowników w bezpiecznym środowisku.
- Środki dotyczące ograniczenia dostępu zostały opracowane w celu zapobieżenia przypadkowej utracie danych, a nie w celu zapobieżenia nieautoryzowanemu dostępowi do danych. (Np. system uprawnień do plików w HDFS zapobiega przed przypadkowym wymazaniem całego całego systemu plików z powdu błędu w programie, lub błędnie wpisanie hadoop fs -rmr /)
- Nie zapobiega złośliwemu użytkownikowi przejęcia tożsamości roota, aby uzyskać dostęp dostęp lub usunąć wszelkie dane w klastrze
- **Hadoop obsługuje 2 mechanizmy uwierzytelniania: prosty, kerberos**.
- **Mechanizm prosty**, który jest domyślny, wykorzystuje UID klienta w celu określenia nazwy użytkownika, którą przekazuje Hadoopowi. W tym trybie, serwery Hadoop w pełni ufają swoim klientom. Ta domyślna wartość jest wystarczająca dla klastrów, w których każdy użytkownik, który może uzyskać dostęp do klastra, w pełni wiarygodny

Zagrożenia bezpieczeństwa:

- Usługi Hadoop nie uwierzytelniają użytkowników ani innych usług. W związku z tym Hadoop jest narażony na następujące **zagrożenia dla bezpieczeństwa**:
    - Użytkownik ma dostęp do klastra HDFS lub MapReduce jak każdy inny użytkownik. Uniemożliwia to egzekwowanie kontroli dostępu w niepewnym środowisku. Np. kontrola uprawnień do plików w HDFS może być łatwo obchodzona.
    - Napastnik może zamaskować się jako usługa Hadoop. Np. kod użytkownika uruchomiony na klastrze MapReduce może się zarejestrować jako nowy TaskTracker.

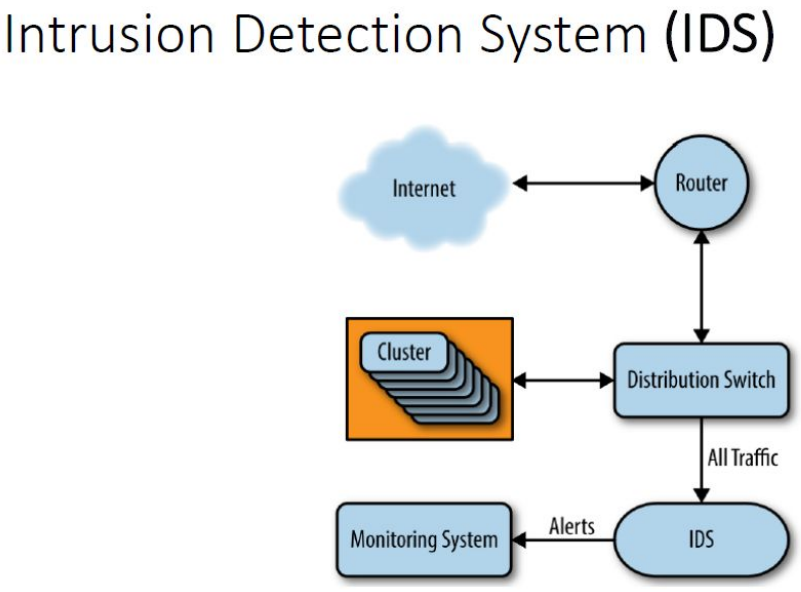
- DataNodes nie egzekwuje żadnej kontroli dostępu do swoich bloków danych. Dzięki temu nieautoryzowany klient może odczytać blok danych tak długo, jak długo może podać swój identyfikator bloku. Każdy może również zapisywać dowolne bloki danych do DataNodes.

Teoria bezpieczeństwa informacji - model CIA:

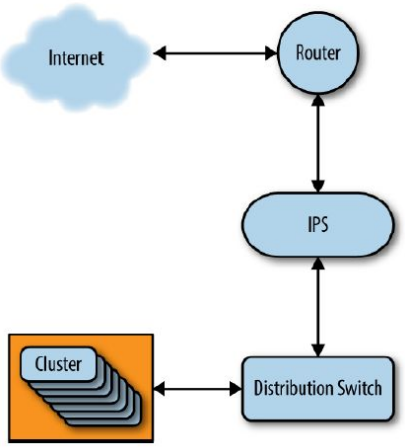
- Poufność(C) - jest zasadą bezpieczeństwa skupiającą się na założeniu, że informacje są dostępne tylko dla zamierzonych odbiorców.
- Integralność(I) - oznacza utrzymywanie i zapewnienie dokładności i kompletności danych w całym cyklu życia. Oznacza to, że dane nie mogą być modyfikowane w sposób nieautoryzowany lub niezauważalny.
- Dostępność(A) - System o wysokiej dostępności mają na celu utrzymanie stałej dostępności, zapobiegając przerwą w świadczeniu usług spowodowanym przerwami w dostawie prądu, awariami sprzętu i modernizacją systemu.

AAA - Authentication, Authorization, Accounting

- Identyfikacja - to stwierdzenie, kto jest kimś lub czym coś jest. Zazwyczaj zgłoszenie jest w formie nazwy użytkownika.
- Uwierzytelnienie - jest aktem weryfikacji stwierdzenia tożsamości. Wprowadzając prawidłowe hasło, użytkownik dostarcza dowód, że jest osobą, do której należy jego nazwa użytkownika.
- Autoryzacja - po udanej identyfikacji i uwierzytelnieniu osoby, programu lub komputera należy określić, do jakich zasobów informacyjnych mają dostęp i jakie działania będą mogły być wykonywane.
- Audytowanie - jest mechanizmem umożliwiającym śledzenie tego, co użytkownicy i usługi robią w klastrze. Jest to krytyczny element systemu bezpieczeństwa, ponieważ bez niego mogą wystąpić naruszenia bezpieczeństwa, których nikt nie zauważy, dostarcza zapis tego, co się stało.



Intrusion Prevention Systems (IPS)



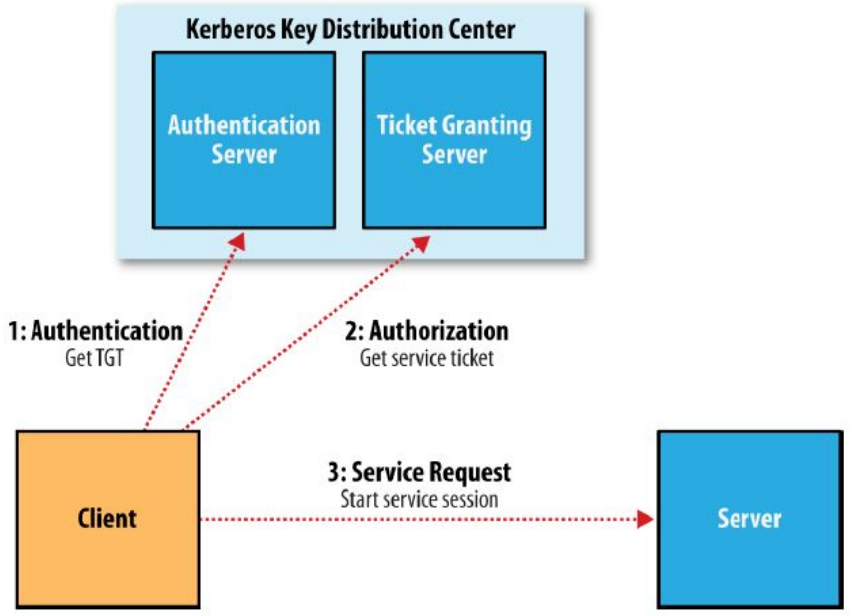
Wymagania bezpieczeństwa:

- Użytkownicy mogą uzyskać dostęp do plików HDFS tylko wtedy, gdy mają do nich uprawnienia.
- Użytkownicy mogą uzyskać dostęp lub modyfikować tylko swoje zadania MapReduce.
- Uwierzytelnianie użytkownika i usługi, aby zapobiec nieautoryzowanemu dostępowi do NameNodes, DataNodes, JobTrackers lub TaskTrackers.
- Usługa do obsługi wzajemnego uwierzytelniania w celu uniemożliwienia nieautoryzowanym usługom dołączania do HDFS lub MapReduce klastra.
- Nabywanie i korzystanie z uprawnień Kerberosa będzie przejrzyste dla użytkownika i aplikacji, pod warunkiem, że system operacyjny nabył Kerberos Ticket Granting Tickets(TGT) dla użytkownika przy logowaniu.
- Degradacja wydajności GridMix powinna wynosić nie więcej niż 3%

Kerberos - the Nerwork Authentication Protocol:

- Kerberos mechanizm uwierzytelniania opracowany w Massachusetts Institute of Technology(MIT). Kerberos stał się faktycznie standardem silnego uwierzytelniania dla systemów komputerowych dużych i małych
- Został on zaprojektowany w celu zapewnienia silnego uwierzytelnienia dla aplikacji klienckich/serwerowych poprzez zastosowanie kryptografii tajnych kluczy.
- Kerberos wykorzystuje symetryczne klucze, które są o rząd wielkości szybsze niż operacje klucza publicznego używane przez SSL.
- Prostsze zarządzanie użytkownikami. Np. cofnięcie uprawnień użytkownika może być dokonane poprzez proste usunięcie użytkownika z centralnie zarządzanego KDC(centrum dystrybucji kluczy) Kerberos. Natomiast w przypadku SSL należy wygenerować nową listę cofnięć certyfikatów i rozesłać ją na wszystkie serwery.

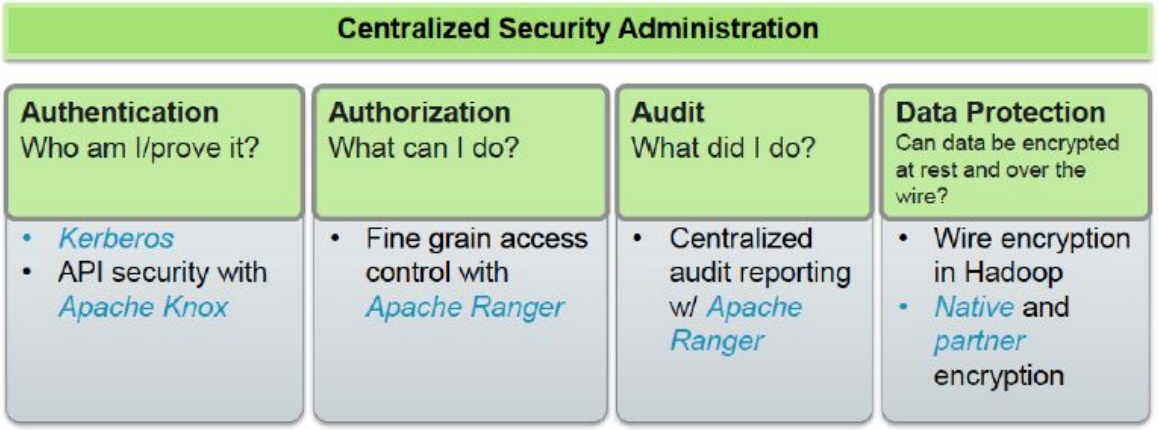
- Protokół wymiany kluczy Kerberos'a:



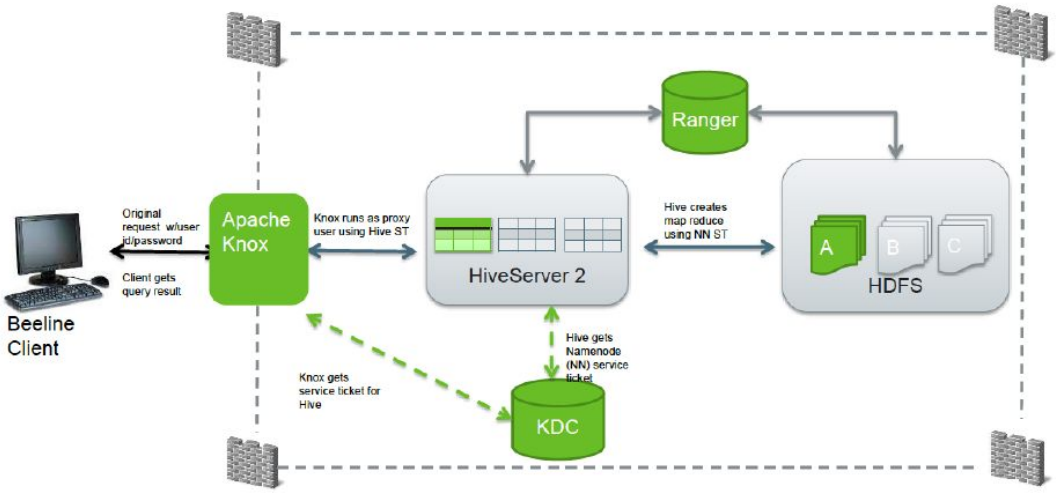
Założenia bezpieczeństwa:

- Hadoop nie wydaje poświadczeń użytkownika ani nie tworzy kont dla użytkowników. Hadoop zależy od danych logowania użytkownika zewnętrznego(np. login systemu operacyjnego, dane logowania Kerberos). Oczekuje się, że użytkownicy uzyskają te poświadczenia od Kerberos podczas logowania do systemu operacyjnego. Usługi Hadoop powinny również być skonfigurowane przy użyciu odpowiednich poświadczeń, w zależności od konfiguracji klastra, w celu wzajemnego uwierzytelnienia.
- Każdy klaster jest konfigurowany niezaleznie. Aby uzyskać dostęp do wielu klastrów, klient musi uwierzytelnić się w każdym klastrze osobno. Jednak pojedyńcze logowanie, które uzyskuje token Kerberos, będzie działać na wszystkich klastrach.
- Użytkownicy nie będą mieli dostępu do kont root w klastrze lub na komputerach używanych do uruchamiania zadań
- Komunikacja HDFS i MapReduce nie będzie działać w niezaufanych sieciach.
- Zadanie Hadoop będzie działać nie dłużej niż 7 dni(konfigurowalne) w klastrze MapReduce lub dostęp do HDFS z zadania zakończy się niepowodzeniem.
- Bilety Kerberos nie będą przechowywane w zadaniach MapReduce i nie będą dostępne dla zadań zadania. Dostęp do HDFS będzie autoryzowany za pomocą tokenów

Technologie bezpieczeństwa w Hadoop:



- **Apache Ranger:**
    - Kompleksowo zapewnia bezpieczeństwo w całym ekosystemie Apache Hadoop
    - Firmy mogą wykonywać wiele zadań, w środowisku wielu najemców
    - Bezpieczeństwo danych w ramach Hadoop musi obsługiwać przypadki wielokrotnego wykorzystania dostępu do danych, jednocześnie zapewniając ramy dla centralnej administracji politykami bezpieczeństwa i monitorowania dostępu użytkowników.
    - Scentralizowana administracja bezpieczeństwa w celu zarządzania wszystkimi zadaniami związanymi z bezpieczeństwem w centralnym UI lub przy użyciu interfejsów REST API.
    - Drobiazgowe uprawnienia do wykonywania określonych działań i/lub operacji za pomocą komponentu/narzędzia Hadoop i zarządzanie nimi za pomocą centralnego narzędzia administracyjnego
    - Jedna metoda autoryzacji we wszystkich komponentach Hadoopa
    - Wsparcie dla różnych metod autoryzacji - kontrola dostępu oparta na rolach, kontrola dostępu oparta na atrybutach itp.
    - Centralizacja kontroli dostępu użytkowników i działań administracyjnych (związanych z bezpieczeństwem) we wszystkich komponentach Hadoopa.
- **Apache Knox:**
    - Bramka API Knox została zaprojektowana jako odwrotny serwer pośredniczący (proxy) na potrzeby zapewnienia bezpieczeństwa bezpieczeństwa.
    - Egzekwowanie polityki sięga od uwierzytelniania/federacji, autoryzacji, audytu, rozsyłki, mapowania
    - Klaster jest zdefiniowany w ramach deskryptora topologii i pozwala bramkom Knox na routing i translację pomiędzy adresami URL użytkowników i wewnętrznymi zasobami klastra.
    - Każdy klaster Apache Hadoop, który jest chroniony przez Knoxa, ma swój zestaw REST API reprezentowany przez specyficzną ścieżkę kontekstową aplikacji. Pozwala to Knox zarówno chronić wiele klastrów, jak i prezentować konsumentowi REST API z jednym punktem dostępu do wszystkich wymaganych usług, w obrębie wielu klastrów.



12. **Apache Spark: charakterystyka, zastosowanie, zasada działania, podstawowe moduły, obsługiwane języki programowania, różnice względem Hadoop MapReduce, RDD (transformacje, akcje), nowe API oparte o DataSet i DataFrame, przykłady zastosowania Spark w obszarze uczenia maszynowego (Machine Learning).**

Apache Spark: to platforma programistyczna do obliczeń rozproszonych. Spark jest wydajnym i skalowalnym silnikiem przetwarzania danych o dużych rozmiarach. Spark został opracowany, aby zapewnić większą prędkość oraz łatwiejszą obsługę niż Hadoop MapReduce**.**

charakterystyka:

1. Skalowalny: dzieli obliczenia na zadania i rozprasza je po węzłach klastra. Potrafi sam zarządzać klastrem danych (ma własny Cluster Manager), jednak potrafi także pracować ściśle z Hadoop i używać Yarn lub inne zgodne menedżery (np. Mesos).
2. Przetwarzanie w pamięci: przetwarzanie w pamięci jest szybsze w porównaniu do Hadoop, ponieważ nie ma czasu poświęcanego na przenoszenie danych z/na dysk. Spark jest prawie 100 razy szybszy niż MapReduce.
3. Przetwarzanie strumieniowe: Spark obsługuje przetwarzanie strumieniowe, które obejmuje ciągłe wprowadzanie i wysyłanie danych.
4. Lazy evaluation (leniwe wartościowanie): Spark rozpocznie wartościowanie/ocenę tylko wtedy gdy jest to absolutnie potrzebne.
5. Obsługuje wiele języków: Spark posiada wbudowane interfejsy API w Javie, Scali oraz Pythona.

Zaawansowana analityka: Spark oprócz obsługi operacji ‘map’ oraz ‘reduce’ obsługuje również zapytania SQL, dane strumieniowe, uczenie maszynowe i algorytmy graficzne.

zastosowanie:

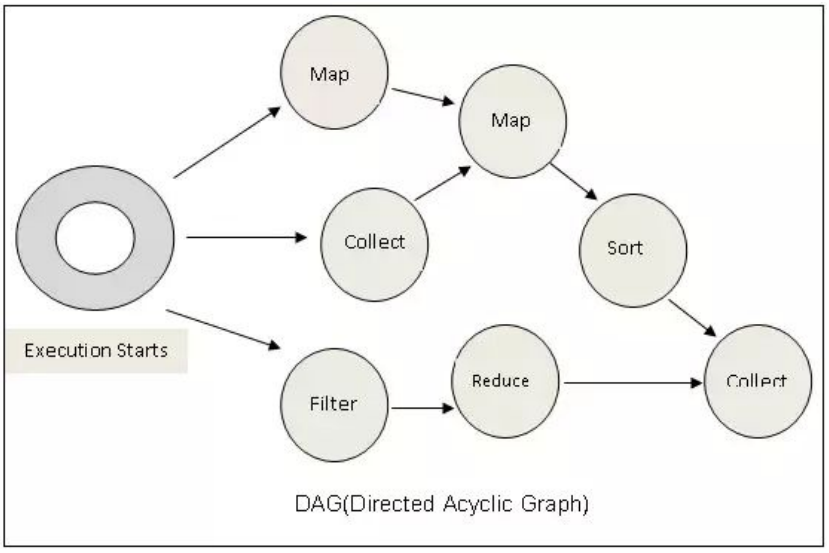
Handel elektroniczny: (ebay, alibaba) - analiza duzych zbiorów danych. Zapewnienie rabatów klientom w opraciu o ich wczesniejsze zakupy.

Opieka zdrowotna: (MyFitnessPal) - zapewnienie lepszych usług klientom

Media i rozrywka: (Netflix) - wyświetlanie odpowiednich reklam użytkownikom na podstawie ich poprzedniej aktywności.

zasada działania:
Wszystko zaczyna się od stworzenia obiektu obsługiwanego w Spark - RDD (Resilient Distributed Dataset).

Kiedy wykonasz operację na nowym RDD, który daje nam końcowy wynik, kontekst Spark przekazuje program do sterownika. Sterownik tworzy dla programu DAG (Directed Acyclic Graph - ukierunkowany graf acykliczny). Po utworzeniu DAG sterownik dzieli go na kilka etapów. Etapy następnie są dzielone na mniejsze zadania, a każde z zadan przekazywane jest wykonawcom do wykonania. (Lazy evaluation, DAG tworzony tylko kiedy jest potrzebny)



Zasada działania z wykładów:

Główną ideą spark jest **RDD (Resilient Distributed Dataset)** - model rozproszonych danych. Metody obliczeniowe przyjmują na wejściu obiekty RDD i wynikiem ich przetwarzania jest zredukowany obiekt RDD.
Źródłem danych dla RDD mogą być:

- pliki z tradycyjnych lub rozproszonych systemów plików,
- pliki ustrukturalizowane (Json, CSV itp.)
- zapytania Hive,
- zewnętrzne usługi (np. HBase, ElasticSearch, Cassandra, JDBC itp.)

Podstawowe moduły

- **Spark streaming** - Analiza strumieniowa danych - np. zdarzeń z systemu lub logów.
- **Spark SQL** - Interfejsy do baz danych
- **MLLib** - Metody zaawansowanej statystyki i uczenia maszynowego
- **GraphX** - Wsparcie dla modeli grafowych np. sieci społecznościowe

  obsługiwane języki programowania

  java, python, scala

  różnice względem Hadoop MapReduce:

MapReduce	Spark
Przetwarzanie wsadowe	Przetwarzanie wsadowe i strumieniowe
Wolniejsze (spowodowane I/O z dysku)	szybsze (działa na RAM)
Niełatwy w użyciu	Przejrzyste API
Działa wolniej ponieważ obsługuje różne formaty danych	Szybsze bo: Bazuje na RDD
Wiecej linii kodu	mniej linii kodu

RDD (transformacje, akcje):
RDD - model rozproszonych danych. Metody obliczeniowe przyjmują na wejściu obiekty RDD i wynikiem ich przetwarzania jest zredukowany obiekt RDD.

Typy operacji wykonywanych na RDD:

- **Transformacje** - funkcje operujące na danych obiektu RDD i tworzące nowy obiekt RDD będący wynikiem ich działania (map, flatMap, filter, union, intersection, distinct, groupByKey, reduceByKey, sortByKey, join, cartesian).
- **Akcje** - funkcje operujące na elementach obiektu RDD, mogą zwracać wartość uzyskaną w trakcie ich wykonania. Typ zwracanej wartości może być różny, w zależności od rodzaju akcji (reduce, collect, count, first, take, countByValue, countByKey).

Nowe API oparte o DataSet i DataFrame

- **DataFrame** - to kolekcje danych zorganizowane w kolumny (tak jak tabele w bazach danych). Każda kolumna ma zdefiniowaną nazwę oraz typ.

- **DataSet** - to rozwinięcie idei DataFrame o pojęcie obiektowości. Są to kolekcje, w których schemat danych zdefiniowany jest przez klasę JVM.

Przykłady zastosowania Spark w obszarze uczenia maszynowego (Machine Learning)

Klasyfikacja, regresja, drzewa decyzyjne, rekomendacje, klastering, analiza przetrwania, rozproszona algebra liniowa, statystyka.

13. Sposoby oceny jakości próbkowania sieci

Próbkowanie możemy uznać za efektywne jeśli:
- Uzyskana próbka zachowuje określone właściwości sieci.
- Analizy własności próbki sieci, na przykład analiza centralności ścieżek, daje wyniki podobne do analiz własności kompletnej sieci
- Uzyskana próbka jest znacznie mniejsza niż sieć pierwotna

Kryterium oceny I(Zachowanie właściwości): w jaki sposób próbka aproksymuje własności całej sieci
Kryterium oceny II(Własności predykcyjne): czy model predykcyjny uczony z wykorzystaniem próbki umożliwi predykcję nieznanych własności:
- Predykcja typu węzła: przewidywanie typu węzła w sieci kompletnej na podstawie danych z próbki
- Przewidywanie brakujących relacji: predykcja i odzyskiwanie brakujących linków
- Funkcje:

• in/out deg; avg in/out deg)
• Jaccard's Coefficient
• P(type(v)|Gs)= $\frac{\#type(v)=t \forall v \in N(n)}{|N(n)|}$
• fRPnode = $\prod_{i \in N(n)} \frac{1}{Z} RP(type(i)|type(v)=t)P(type(v)=t)$
• fRPpath = $\sum_{p \in Path(s,t)} \prod_{(p_1,p_2) \in p} P(type(p_2)|type(p_1))$

14. Różnice między agregacją danych z sieci a próbkowaniem

Próbkowanie	Agregacja
Informacja o węzłach/ linkach jest pozyskana dopiero po pobraniu próbki	Znana jest cała struktura sieci apriori

Wymaga strategii eksploracji sieci i stopniowego powiększania próbki	x
Cel : stopniowa identyfikacja małego zbioru przedstawiciel i węzłów i powiązań ze struktury sieciowej , przy posiadanej niewielkiej wiedzy o całej sieci	Cel: zagregowane miary, które umożliwią opis własności sieci na poziomie ogólnym przy jak najmniejszej utracie informacji szczegółowych.

// tabelka od zaocznych

Pobieranie próbek:

- Informacja o węzłach/linkach jest pozyskana dopiero po pobraniu próbki
- Wymaga strategii eksploracji sieci i stopniowego powiększania próbki
- Celem jest stopniowa identyfikacja małego zbioru przedstawicieli węzłów i powiązań ze struktury sieciowej, przy posiadanej niewielkiej wiedzy o całej sieci.

Agregacja danych:

- Znana jest cała struktura sieci apriori
- Celem są zagregowane miary, które umożliwią opis własności sieci na poziomie ogólnym przy jak najmniejszej utracie informacji szczegółowych.

15. Główne strategie próbkowania sieci homogenicznych

Trzy główne strategie:

- ***Wybór węzła****:*
  - Losowy wybór węzła(losowy zbiór węzłów)
  - Wybór na podstawie stopnia wierzchołka(Prawdopodobieństwo proporcjonalne do jego degeee - zakładamy, że degree jest znane)
  - PageRank sampling (Prawdopodobieństwo wyboru węzła jest proporcjonalne do wartości jego miary PageRank-zakładając, że jest znana)
- ***Wybór krawędzi:***
  - Random Edge Sampling(RE) - krawędzie wybieramy losowo, a następnie włączone są powiązane nimi węzły
  - Random Node-Edge Sampling(RNE) - wybieramy węzły a następnie powiązane z nimi krawędzie
  - Hybrid Sampling - z prawdopodobieństwem p realizowany jest RE sampling, a z prawdopodobieństwem 1-p RNE sampling.
  - Induced Edge Sampling - Krok 1: jednolity wybór krawędzi(a w konsekwencji węzłów) przez kilka rund. Krok 2: dodawane są krawędzie, które są powiązane z wybranymi węzłami
  - Frontier sampling - Krok 0: Losowo wybieraj zestaw węzłów L jako seeds. Krok 1: Wybierz element u z L przy użyciu degree based sampling. Krok 2: Wybierz krawędzie węzła u(u,v). Krok 3: Zastąp “u” przez “v” w zbiorze i dodaj(u,v) do sekwencji próbkowanych węzłów. Powtórz kroki 1 do 3.
- ***Sampling w procesie eksploracji***

- Random walk - węzeł z następnego przeskoku jest wybierany jednolicie wśród sąsiadów bieżącego węzła
- Random walk z restartem - wybór węzła, random walk i ponowne uruchomienie
- Random jump - Podobnie jak random walk, ale z dodatkowo z prawdopodobieństwem p następują przeskoki do innych losowo wybranych węzłów sieci
- Forest fire - Wybór węzła u. Losowe generowanie liczby “z”(<= liczba linków węzła u) i selekcja z linków jeszcze nie odwiedzonych. Krok wykonywany rekursywnie dla wszystkich nowo dodanych węzłów
- Ego-centric exploration&sampling(ECE) - Zmodyfikowany random walk z przypisanymi prawdopodobieństwami selekcji uzależnionymi od właściwości węzła.
- Depth First/Breadth-First - próbkowanie sąsiadów najczęściej odwiedzanych węzłów lub ostatnio odwiedzonych.
- Sample Edge Count - przekierowanie do sąsiada z najwyższym degree i kontynuacja od niego
- Expansion sampling - konstruowanie próbek tak by maksymalizować ekspansję.

16. Różnie między Random Walk a Snowball Sampling

Random Walk - Technika losowa. Węzeł z następnego przeskoku jest wybierany jednolicie wśród sąsiadów bieżącego węzła.

Snowball Sampling - to nielosowa technika doboru węzła(respondentów) do próby badanej. Np. po zakończeniu każdego kolejnego wywiadu ankieter prosi respondenta o wskazanie znajomej osoby, z którą mógłby również przeprowadzić wywiad. Przydatna w badaniach grup społecznych.

17. Główne cechy języka Cypher.
1. (Nawiasy okrągłe) oznaczają węzły
2. [Nawiasy kwadratowe] określają relacje
3. {Nawiasy klamrowe} mówią nam o właściwościach.

4. Deklaratywny grafowy język zapytań
5. Zapytania i aktualizacje
6. Inspirowany przez SQL i SPARQL (wzorce)

http://adam.wroclaw.pl/2014/09/grafowa-baza-danych-neo4j/

18. **Omówić korzyści wynikające z zastosowania grafowych baz danych do przeszukiwania połączeń lotniczych w porównaniu do relacyjnych baz danych.**

Korzyści wynikają z przedstawienia problemu w postaci grafu: atrybuty mają węzły oraz krawędzie między węzłami. Wykorzystując w pełni strukturę grafu możemy prościej pisać złożone zapytania do bazy.

Korzyści

- prosty model
- mniej skompikowane, intuicyjne zapytania
    - ten przypadek wymaga wielu joinów i podzapytań
- szybsze wyszukiwanie danych

Grafy	Relacyjne bazy danych
połączenia są opisane jako ciąg “przystanków” a-> b -> c	połączenia do każdego lotniska opisane jako osobne rekordy: a->b , b->a
mniej skompikowane zapytania, bardzo czytelne	skomplikowane zapytania
prosty model	dużo joinów i podzapytań
	wiersze wypełnione niewiele mówiącymi liczbami - identyfikatorami
	tworzenie złożonych baz jest bardzo czasochłonne
	dane muszą być unikalne lub pasować do wzorca

“Potencjalne” pytania

- na labach Jankowski rzucił “aż się prosi o pytanie jakie są technologie i algorytmy próbkowania”
    - Random Walk
    - Random Walk z restartem
    - Random jump
    - Forest fire
    - Ego-centric exploration & sampling
    - Depth First/ Breadth-First
    - Sample Edge Count
    - Expansion Samplingi
- z facebooka : czym się różni Cypher od sqla

- Cypher jest zbudowany na podstawie podstawowych konceptów i klauzul SQL’a z dodatkiem funkcjonalności grafowych
- Cypher jest bardziej czytelny
- SQL jest dobry dla relacyjnych baz danych, ale nie dla dużych, złożonych danych
-

Czym są duże zbiory danych: charakterystyka, obszary zastosowań.

Termin „**duże zbiory danych**” odnosi się do wszystkich danych biznesowych o dużych objętościach (potencjalnie wiele terabajtów), które muszą być przechowywane, przetwarzane i analizowane w efektywny i ekonomiczny sposób, i które mogą wymagać udostępniania online. Oprócz coraz większej ilości danych, zbiory te są zazwyczaj również w wysokim stopniu nieustrukturyzowane i posiadają wiele wariantów danych pochodzących z różnych źródeł, w tym danych historycznych.
Przetwarzanie dużych zbiorów danych wymaga wyeliminowania wszystkich uszkodzonych lub powielających się danych oraz przekonwertowania pozostałych do formatu, w którym mogą być wykorzystane do stymulowania zrównoważonego wzrostu biznesu.

Słowa kluczowe Big daty

- ➔ Ilość (volume)
- ➔ Szybkość (velocity)
- ➔ Różnorodność (variety)
- ➔ Zmienność (variability)
- ➔ Złożoność (complexity)

Obszary zastosowań:

- ➔ Banki
- ➔ Produkcja
- ➔ Handel detaliczny
- ➔ Ochrona zdrowia
- ➔ Sektor publiczny
- ➔ Edukacja

Hadoop: charakterystyka, zastosowania, elementy składowe, zapewnienie niezawodności działania

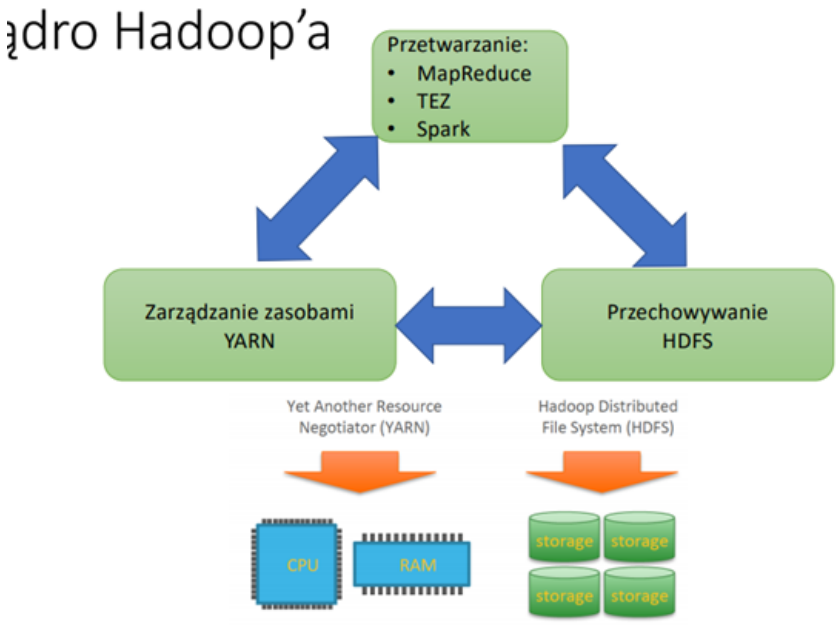
Hadoop - otwarta platforma programistyczna napisana w języku Java przeznaczona do rozproszonego składowania i przetwarzania wielkich zbiorów danych przy pomocy klastrów komputerowych. Jest jednym z projektów rozwijanych przez fundację Apache.

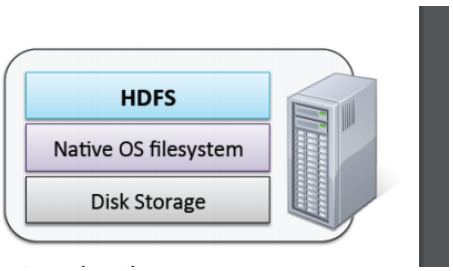
Cechy:

- ➔ Elastyczna struktura danych
- ➔ Skalowalność
- ➔ Złożone przetwarzanie danych
- ➔ Dane są kopiowane do pliku bez żadnej transformacji
- ➔ Mechanizm serializacji jest wykorzystany do odczytu danych aby uzyskać dostęp do kolumn.
- ➔ Nowy typ danych może przyjść w dowolnym momencie i wykorzystany po aktualizacji mechanizmów serializacji.
- ➔ Szybki zapis

Zastosowanie:

- ➔ Extract/Transform/Load (ETL)
- ➔ Eksploracja tekstu - text mining
- ➔ Budowanie indeksów
- ➔ Analizy grafowe
- ➔ Rozpoznawanie wzorców - pattern recogniton
- ➔ Systemy rekomendujące, np. collaboratve filtering
- ➔ Modele predykcyjne
- ➔ Analiza sentymentu
- ➔ Ocena ryzyka





BigData vs. relacyjne bazy danych.

Relacyjne bazy danych:

- ➔ Struktura danych musi być zdefiniowana przed zapisem
- ➔ Muszą być zdefiniowane transformacje danych do wewnętrznej struktury bazy
- ➔ Nowe kolumny muszą być zdefiniowane (dodane) nim dane będą załadowane do bazy.
- ➔ Szybki odczyt
- ➔ Standaryzacja

Big data:

- ➔ Dane są kopiowane do pliku bez żadnej transformacji.
- ➔ Mechanizm serializacji jest wykorzystany do odczytu danych aby uzyskać dostęp do kolumn.
- ➔ Nowy typ danych może przyjść w dowolnym momencie i wykorzystany po aktualizacji mechanizmów serializacji.
- ➔ Szybki zapis
- ➔ Elastyczność

HDFS: charakterystyka, redundancja danych w HDFS, szybkość zapisu i odczytu, niezawodność, zadania Namenode i Datanode.

HDFS

- ➔ jest systemem plików napisanym Javie. W oparciu o Google File system
- ➔ Zapewnia nadmiarową pamięć masową dla ogromnych ilości danych
- ➔ Korzysta z łatwo dostępnych, standardowych w branży komputerów
- ➔ HDFS najlepiej sprawdza się przy "niewielkiej" ilości dużych plików

Redundancja danych:

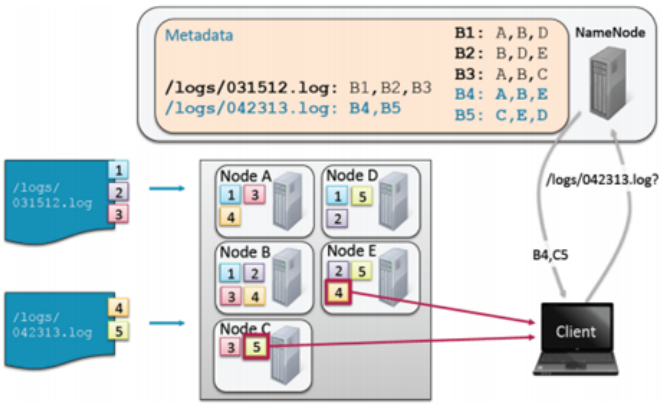
Bloki i metodologia replikacji Hadoop Distributed File System (HDFS) mają dwa kluczowe pojęcia, tj. „Rozmiar bloku” i „współczynnik replikacji”. Każdy plik, który trafia do HDFS, jest podzielony na kilka części lub „bloków”.
Liczba bloków zależy od maksymalnego przydzielonego rozmiaru bloku, zwykle 128 MB. Po utworzeniu bloków są one replikowane w klastrze HDFS. Liczba replik jest określana przez współczynnik replikacji (RF), zwykle konfigurowany jako 3 (1 oryginał i 2 kopie).
Ta redundancja pomaga w budowaniu odporności i odporności na błędy, tzn. Gdy jeden z bloków zawiedzie, mamy kolejne 2, z których można bezpiecznie przywrócić dane.

Szybkość zapisu i odczytu:

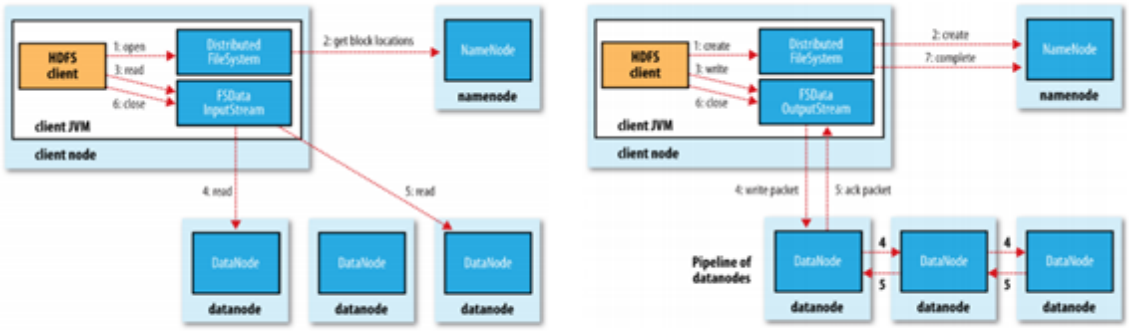
zapis:
hdfs , a dokładnie to NameNode buforuje wszystkie nazwy plików i adresy bloków w pamięci (Po ludzku: wie co i gdzie leży). Dzięki takiemu mechanizmowi HDFS jest szybki. Duża ilość plików oznacza dużą ilość metadanych, którą musi ogarnąć JVM. Nie edytować plików już zapisanych w HDFS, ale możemy dołączyć dane, ponownie otwierając plik

Odczyt:
HDFS jest zoptymalizowany do dużych, strumieniowych odczytów plików. Zamiast losowych odczytów

Zapis i odczyt danych w HDFS



HDFS – odczyt i zapis



Zadania Namenode i Datanode:

Namenode:
NameNode zapewnia uprawnienia, dzięki czemu klient może łatwo odczytywać i zapisywać bloki danych do/z odpowiednich węzłów danych.
Przechowuje drzewo katalogów wszystkich plików w systemie plików i śledzi, gdzie w klastrze przechowywane są dane plików. Sam nie przechowuje danych tych plików.

- ➔ Demon NameNode musi być bez przerwy uruchomiony (W przypadku zatrzymania się węzła NameNode, klaster staje się niedostępny)

Datanode:
DataNode przechowuje dane w [HadoopFileSystem]. Funkcjonalny system plików ma więcej niż jeden węzeł DataNode, w którym są replikowane dane. Podczas uruchamiania DataNode łączy się z NameNode; kręci się, aż pojawi się ta usługa. Następnie odpowiada na żądania z NameNode dotyczące operacji na systemie plików.
Aplikacje klienckie mogą komunikować się bezpośrednio z DataNode, gdy NameNode poda lokalizację danych.

YARN: charakterystyka, tryby pracy.

YARN - Yet Another Resource Negotiator. Jest warstwą przetwarzania Hadoopa, która zawiera:

- ➔ Zarządcę zasobów (ang. resource manager)
- ➔ Harmonogram pracy (ang. job scheduler)

YARN pozwala na pracę wielu silników przetwarzających dane na jednym klastrze Hadoop:

- ➔ Programy wsadowe (np. Spark, MapReduce)
- ➔ Interaktywny SQL (np. Impala)
- ➔ Zaawansowane funkcje analityczne (np. Spark, Impala)
- ➔ Streaming (np. SparkStreaming).

Tryby Pracy(?):

- ➔ Resource Manager (RM)
  - ◆ Działa na węźle głównym
  - ◆ Globalny harmonogram zasobów
  - ◆ Arbitrażuje zasoby systemowe pomiędzy konkurencyjnymi aplikacjami
  - ◆ Ma wbudowany harmonogram do obsługi różnych algorytmów
- ➔ Node Manager (NM)
  - ◆ Działa na węzłach podrzędnych (slave)
  - ◆ Komunikuje się z RM
- ➔ Kontenery
  - ◆ Utworzone przez RM na żądanie.
  - ◆ Przydzielenie określonej ilości zasobów (pamięć, CPU) na węźle podrzędnym (slave)
  - ◆ Aplikacje wykonywane są w jednym lub kilku kontenerach
- ➔ Application Master (AM)
  - ◆ Jeden na aplikację.
  - ◆ Specyficzny dla aplikacji/usługi
  - ◆ działa w kontenerze
  - ◆ Może żądać więcej kontenerów do realizacji zadań aplikacji

Paradygmat MapReduce: charakterystyka, zasada działania, możliwości i ograniczenia.

MapReduce jest metodą rozłożenia zadań na wiele węzłów. Każdy fragment jest przetwarzany równolegle między węzłami w klastrze.

Zadania:
Maper:
- ➔ Każde zadanie mapy (zazwyczaj) działa na pojedynczym bloku HDFS.
- ➔ Zadania mapowania (zwykle) wykonywane są na węźle, w którym blok jest przechowywany

Sortuj i przetasuj (ang. Shuffle and Sort):
- ➔ Sortuje i konsoliduje dane pośrednie ze wszystkich maperów.
- ➔ Następuje po zakończeniu zadań związanych z mapowaniem i przed rozpoczęciem redukcji zadań.

Reduktor:
- ➔ Pracuje na danych pośrednich przesyłanych w sposób losowy/sortowanych (wyjście zadania mapowania).
- ➔ Generuje wynik końcowy

Zasada działania:
Maper pobiera każdy wiersz z tekstu wejściowego jako dane wejściowe i dzieli go na słowa.
Emituje parę klucz/wartość za każdym razem, gdy wyraz wypada wyraz, po którym następuje 1. Dane wyjściowe są sortowane przed wysłaniem go do programu do redukcji.
Zmniejszenie sumuje te pojedyncze zliczenia dla każdego wyrazu i emituje pojedynczą parę klucz/wartość zawierającą wyraz, a następnie sumę jego wystąpień. (więcej na slajdach, które będą dołączone)

Możliwości i ograniczenia:
Ze względu na możliwość wystąpienia awarii wynik pracy węzła jest wysyłany dopiero po zakończeniu wszystkich przydzielonych mu obliczeń. Pozwala to powtórzyć daną partię obliczeń bez ryzyka, że częściowe wyniki zostaną uwzględnione w wynikach dwu lub wielokrotnie. Domyślnie działa tylko jeden reduktor.
Może to powodować istotne problemy, jeśli mamy dużo danych pośrednich. Węzeł, na którym pracuje reduktor, może nie mieć wystarczająco dużo miejsca na dysku, aby pomieścić wszystkie dane pośrednie. Redukcja zajmie dużo czasu.

HIVE: charakterystyka, zastosowania, możliwości i ograniczenia HiveQL, Sqoop.

Charakterystyka:
HIVE jest zbliżonym do SQL-u interfejsem wykorzystującym Hadoop MapReduce, aby uwolnić użytkowników od zajmowania się niskopoziomowymi szczegółami implementacji równoległych zadań przetwarzania wsadowego.

Hive koncentruje się głównie na ekstrakcji - transformacji - ładowaniu (ETL) oraz przetwarzaniu wsadowym (w partiach):

- ➔ odczytywanie ogromnych ilości danych,
- ➔ dokonywanie przekształceń tych danych (np. przemieszanie danych, konsolidacja, agregowanie)
- ➔ załadowanie danych wyjściowych do innych systemów, które są wykorzystywane do dalszej analizy

TEZ: charakterystyka, mechanizmy które pozwoliły na przyspieszenie pracy Hive: ORC, LLAP, wektoryzacja, DAG – skierowane grafy acykliczne.

Podobnie jak Spark, Apache Też to platforma typu open source do przetwarzania dużych zbiorów danych oparta na technologii MapReduce. Zarówno Spark, jak i Też oferują silnik wykonawczy, który jest w stanie używać ukierunkowanych wykresów acyklicznych (DAG) do przetwarzania bardzo dużych ilości danych.

Też uogólnia paradygmat MapReduce, traktując obliczenia jako DAGs.

Zadania MapReduce łączą się w jedno zadanie, które jest traktowane jako węzeł w DAG, wymuszając współbieżność i serializację.wykonuje lepiej operacje MapReduce (ok. 10x szybciej)

Mechanizmy które pozwoliły na przyspieszenie pracy Hive:

ORC (Optimized Record Columnar):

- ➔ W przypadku tabeli przechowywanej w pliku ORC, jest ona najpierw dzielona poziomo na wiele segmentów. Następnie, w segmencie, wartości danych zapisywane są w kolumnach, jedna po drugiej**.**
- ➔ Wszystkie kolumny w segmencie są zapisywane w tym samym pliku. Ponadto, aby dostosować się do różnych wzorców zapytań, zwłaszcza zapytań ad hoc, plik ORC nie umieszcza kolumn w grupach kolumn.
- ➔ Domyślny rozmiar segmentu to 256 MB.
- ➔ Indeksy: wskaźniki pozycji i statystyki
- ➔ W przypadku kolumny o złożonym typie danych (np. Mapa), rozkłada się tę kolumnę na wiele kolumn podrzędnych. Można wybrać opcję dostosowania wielkości segmentu z wielkością blokiem HDFS. Jeden segment zawsze będzie w jednym bloku HDFS.

- ➔ Kompresja:
  - ◆ na poziomie kolumn, np. string można kodować słownikowo.
  - ◆ cały plik: ZLIB, Snappy and LZO.

LLAP(Live Long And Prosper):
Trwałe demony (kontenery) zapytań i inteligentne buforowanie w pamięci, aby dostarczać błyskawicznie szybkie zapytania SQL przy zachowaniu skalowalności.

- ➔ Dzięki trwałym (persistent) demonom zapytań unikamy długiego czasu uruchamiania kodu SQL.
- ➔ Eliminacja przydzielania kontenerów i czasu uruchamiania JVM.
- ➔ Dzieli swoją pamięć podręczną pomiędzy wszystkich użytkowników SQL, maksymalizując wykorzystanie tego zasobu.
- ➔ Dysponuje drobnoziarnistym zarządzaniem zasobami i prewencją, co zapewnia równoczesny dostępu dla wielu użytkowników.
- ➔ Jeśli to możliwe, praca planowana jest na węźle z danymi w pamięci podręcznej, jeśli nie, zostanie wykonana w innym węźle.
- ➔ Jest w 100% kompatybilny z istniejącymi narzędziami Hive SQL i Hive.

Wektoryzacja:
W wektorowym modelu wykonania, zestaw danych jest reprezentowany jako partia wierszy.
- ➔ W partii wierszy wartości danych danej kolumny są reprezentowane jako wektory kolumn. Ilość wierszy w partii jest konfigurowalna i powinna być wybrana tak, aby pasowała do wielkości pamięci podręcznej procesora.
- ➔ Domyślnie ta liczba jest ustawiona na 1024, co zazwyczaj pozwala zmieścić jeden wektor w pamięci podręcznej procesora.
- ➔ Wykonywanie zapytań postępuje poprzez zastosowanie każdego wyrażenia na całym wektorze kolumny

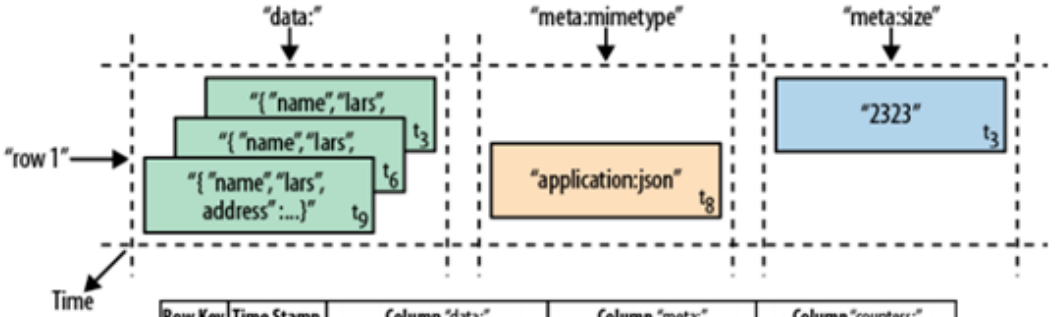
DAG – skierowane grafy acykliczne.

Jest to w informatyce bardzo ważna struktura, łącząca zalety drzew i ogólnych grafów skierowanych. Określa relacje między operatorami (sekwencji lub równoległości zadań), porządek i zależności. Oznacza to, że podczas tworzenia przepływu pracy należy zastanowić się, jak podzielić go na zadania, które mogą być wykonywane niezależnie od siebie.
Następnie możesz połączyć te zadania w logiczną całość, układając je w wykres. Kształt wykresu decyduje o ogólnej logice przepływu pracy. Tworzy to stabilną konstrukcję, ponieważ każde zadanie może być ponawiane wiele razy, jeśli wystąpi błąd.

Hbase: charakterystyka, zasada działania, możliwości i ograniczenia, zadania Master i Regionserver. (tylko studia stacjonarne)

Charakterystyka, możliwości i ograniczenia:
HBase jest rozproszoną bazą danych zbudowaną na bazie HDFS. Bazuje na rozwiązaniu Google Bigtable.

- ➔ HBase jest aplikacją Hadoop przeznaczoną do losowego dostępu do bardzo dużych zbiorów danych w czasie rzeczywistym.
- ➔ Skaluje się ona liniowo poprzez dodawanie węzłów.
- ➔ HBase nie jest relacyjna i nie obsługuje SQL,
- ➔ Jest w stanie zrobić to, czego RDBMS nie jest w stanie zrobić:
  - ◆ -obsługiwać bardzo duże, słabo zapełnione tabele na klastrach wykonanych z typowego sprzętu
- ➔ Dane są przechowywane w tabelach.
- ➔ Tabele składają się z wierszy i kolumn.
- ➔ Zawartość komórki jest nie interpretowaną tablicą bajtów.
- ➔ Kolumny wierszy są pogrupowane w rodziny kolumn.
- ➔ Wszyscy jej członkowie mają wspólny prefiks.
- ➔ Kolumna rodzina i kwalifikator są zawsze oddzielone znakiem dwukropka (:).
- ➔ Rodziny kolumn tabeli muszą być określone z góry jako część definicji schematu tabeli, ale nowi członkowie rodzin kolumn mogą być dodawani na żądanie.
- ➔ Komórki tabeli - przecięcie wiersza i kolumny - są w wersjonowane.



Row Key	Time Stamp	Column "data:"	Column "meta:" "mimetype"	"size"	Column "counters:" "updates"
"row1"	t_3	"{ "name" : "lars", "address" : ...}"		"2323"	"1"
	t_6	"{ "name" : "lars", "address" : ...}"			"2"
	t_8		"application/json"		
	t_9	"{ "name" : "lars", "address" : ...}"			"3"

zadania Master i Regionserver:

Master:

- ➔ Master jest odpowiedzialny za przydzielanie regionów do serwerów regionalnych i używa Apache ZooKeeper, niezawodnej, wysoce dostępnej, stałej i rozproszonej usługi koordynacji, aby ułatwić to zadanie.
- ➔ Serwer master jest również odpowiedzialny za obsługę równoważenia obciążenia regionów na serwerach regionalnych, aby rozładować obciążone serwery i przenieść regiony do mniej obciążonych.
- ➔ Serwer master nie jest częścią faktycznej ścieżki przechowywania lub pobierania danych. Negocjuje on równoważenie obciążenia i utrzymuje stan klastra, ale nigdy nie świadczy żadnych usług związanych z danymi ani dla serwerów regionalnych, ani dla klientów, dlatego w praktyce jest lekko obciążony.
- ➔ Ponadto zajmuje się zmianami schematów i innymi operacjami na metadanych, takimi jak tworzenie tabel i rodzin kolumn.

Regionserver:

RegionServers to procesy oprogramowania (często nazywane demonami), które aktywujesz w celu przechowywania i pobierania danych w HBase (baza danych Hadoop). W środowiskach produkcyjnych każdy RegionServer jest wdrażany w osobnym dedykowanym węźle obliczeniowym. Kiedy zaczynasz korzystać z HBase, tworzysz tabelę, a następnie zaczynasz przechowywać i pobierać dane.

Jednak, w pewnym momencie — i być może dość szybko w przypadkach użycia dużych zbiorów danych — tabela rozrasta się poza konfigurowalny limit. W tym momencie system HBase automatycznie dzieli tabelę i rozdziela obciążenie na inny serwer RegionServer.

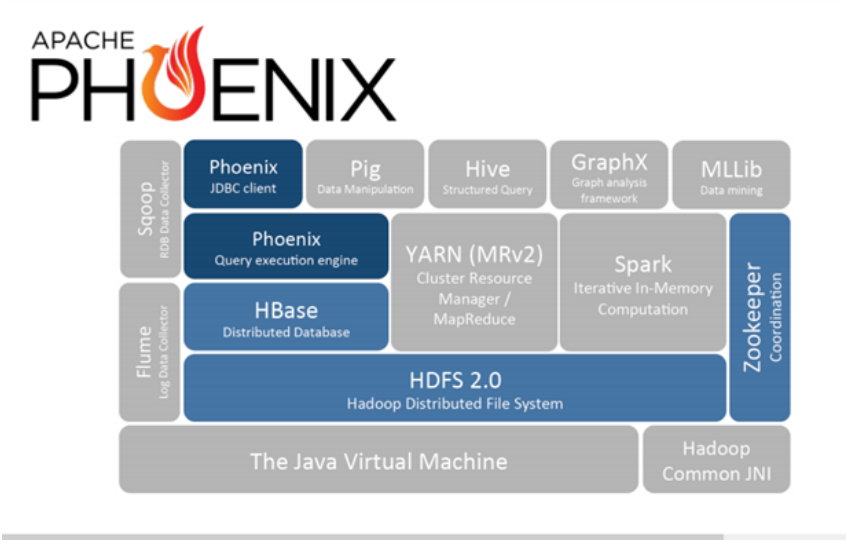
Phoenix: charakterystyka, zasada działania. (tylko studia stacjonarne):

Charakterystyka:

Apache Phoenix umożliwia analizę OLTP i analizę operacyjną w Hadoopie dla aplikacji o niskich opóźnieniach poprzez połączenie zalet obu światów:

- ➔ moc standardowych interfejsów API SQL i JDBC z pełnymi możliwościami transakcji ACID
- ➔ schema-on-read ze świata NoSQL poprzez wykorzystanie HBase jako nośnika danych.

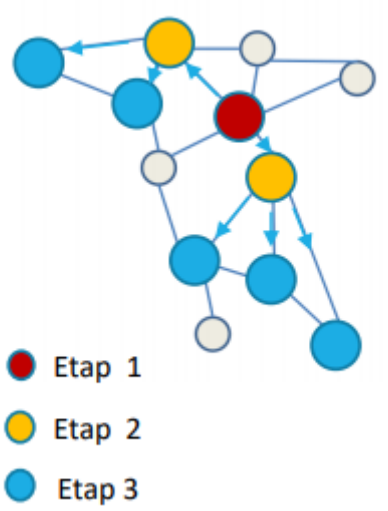
Apache Phoenix jest w pełni zintegrowany z innymi produktami Hadoop, takimi jak Spark, Hive, Pig, Flume i Map Reduce.



Różnice między Random Walk a Snowball Sampling.

Snowball sampling:

Dla wskazanych początkowo węzłów, do próbki włączanych jest n sąsiadów wybieranych losowo. Proces postępuje iteracyjnie.



Random walk:
Dla aktywnego węzła jest wybierany losowo tylko jeden z jego sąsiadów i następuje do niego przejście.