



LEKSIONI 6

Apache Flume

ABSTRACT

Në këtë leksion do të kuptojmë se si për Apache Flume ndihmon në transmetimin(streaming) e të dhënave nga burime të ndryshme.

Një nga arsytet e rëndësishme pse Flume u bë kaq popullor, është se ai mund të integrohet shumë lehtë me Hadoop dhe ngarkojë të dhëna të pastruara, si dhe gjysmë të strukturuara në HDFS. Kjo është arsyeja pse Apache Flume është një pjesë e rëndësishme e Ekosistemit Hadoop.

Në këtë leksion do të shohim:

-

Alba ÇOMO

Menaxhimi i të dhënave të mëdha
Msc Teknologji Informacioni

Përmbajtja

1. Hyrje	2
2. Tranferimi i të dhënave në Hadoop.....	2
2.1 Të dhënat Log/Streaming.....	2
2.2 Problemet me komandën –put në HDFS.....	3
2.3 Zgjidhja për transferimin e të dhënave	3
3. Arkitektura Flume.....	4
3.1 Ngjarja Flume.....	5
3.2 Agjenti Flume.....	5
Komponentë shtesë të një Agjenti Flume.....	6
4. Apache Flume – Rrjedha e të dhënave.....	6
5. Karakteristikat dhe Kufizimet e Apache Flume.....	7
5.1 Karakteristikat	8
5.2 Kufizimet	9
6. Aplikimet e Apache Flume.....	9
7. Apache Flume.....	10
8. Setup -Shembull.....	12

Figurat

Figura 1 Apache Flume	2
Figura 2 Arkitektura Flume	4
Figura 3 Arkitektura Flume	5
Figura 4 Struktura e një ngjarje Flume	5
Figura 5 Struktura e Agjentit Flume.....	6
Figura 6 Rrjedhja e të dhënave në Flume	7
Figura 7 Burimet Flume.....	10
Figura 8 Kanalet Flume	11
Figura 9 Kanalet Flume	12

1. Hyrje

Apache Flume është një mjet/shërbim/mekanizëm data ingestion për mbledhjen, agregimin dhe transportimin e sasive të mëdha të të dhënave streaming, si për shembull skedarët log, ngjarjet, etj nga burime të ndryshme në një data store të qëndrueshëm.

Flume, kryesisht është krijuar për të kopjuar streaming data (të dhëna log) nga web servers të ndryshëm në HDFS. Flume ka një arkiteturë të thjeshtë dhe fleksibël bazuar në rrjedhën e të dhënave streaming. Ai është tolerant ndaj gabimeve dhe siguron mekanizma të besueshëm për tolerancën ndaj gabimeve dhe recuperimin në rast dështimi.

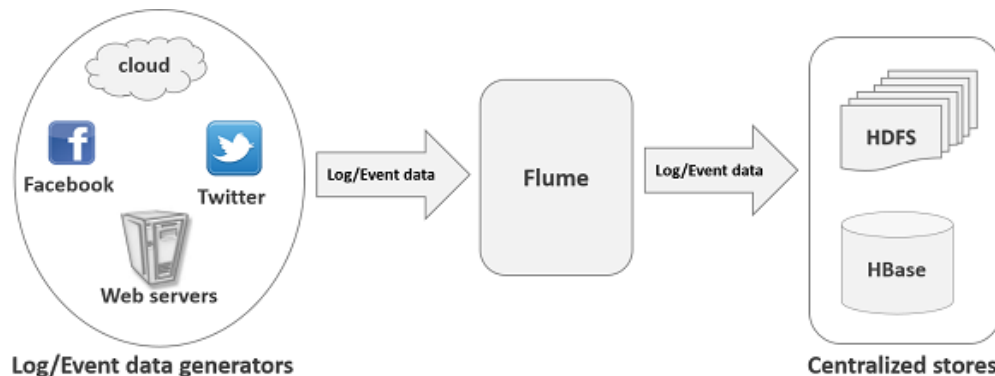


Figura 1 Apache Flume

Supozoni se një aplikacion e-commerce kërkon të analizojë sjelljen e klientëve nga një rajon i caktuar. Për të realizuar këtë, ata duhet të ngarkojnë log-un e të dhënave në Hadoop për të realizuar analizat, kjo realizohet duke përdorur Apache Flume. Flume përdoret për të lëvizur të dhënat log të gjeneruara nga serverat e aplikacionit në HDFS me një shpejtësi të lartë.

2. Tranferimi i të dhënave në Hadoop

Ashtu siç kemi përmendur dhe në leksionet e kaluara, Big Data është një koleksion i madh i bashkësive të të dhënave të cilat nuk mund të përpunohen nga teknikat tradicionale. Të dhënat e mëdha, kur analizohen, japin rezultate të rëndësishme dhe të vlefshme. Hadoop është një framework open-source i cili mundëson ruajtjen dhe përpunimin e të dhënave të mëdha (big Data) në një mjedis të shpërndarë nëpër cluster-a, kompjutera të lidhur të cilët përdorin një model programi të thjeshtë.

2.1 Të dhënat Log/Streaming

Në përgjithësi, shumë nga të dhënat që do të analizohen sigurohen nga burime të ndryshme të dhënash, si për shembull serverat e aplikacioneve, rrjetet sociale, serverat cloud dhe serverat e ndërmarrjeve. Këto të dhëna janë në formën e skedarëve log apo të ngjarjeve.

Skedarët Log – Në përgjithësi, një skedar log është një skedar që ruan/regjistron ngjarjet apo veprimet në një sistem operimi. Për shembull, web serverat regjistrojnë çdo kërkesë që është realizuar në server. Nga këto skedarë ne mund të marrim informacione mbi:



Performancën e aplikacionit dhe të lokalizojmë dështime të ndryshme të software-it apo hardware-it.



Sjelljen e përdoruesve dhe nxjerrjen e njohurive më të mira të biznesit.

Metoda tradicionale për transferimin e të dhënave në sistemet HDFS është përdorimi i komandës `put`. Sfidë më e madhe në përdorimin e të dhënave log është lëvizja e këtyre log-eve të gjeneruara nga servera të ndryshëm në mjedin Hadoop.

2.2 Problemet me komandën `put` në HDFS

Hadoop *File system Shell* ofron komanda për futjen e të dhënave në Hadoop dhe leximin e tyre prej tij. Në mund të insertojmë të dhëna në Hadoop duke përdorur komandën `put` si mëposhtë:

```
hdfs dfs -put /path i skedari i ruajtuar lokalisht /path në HDFS ku do të ruhet skedari
```

Përdorimi i kësaj komande shfaq problemet e mëposhtme:



Duke përdorur komandën **put**, ne mund të transferojmë vetëm një skedar në një kohë ndërsa gjeneruesi e të dhënave gjenerohen të dhëna me një shpejtësi shumë më të lartë. Meqenëse analiza mbi të dhëna të vjetra është më pak e saktë, ne duhet të kemi një zgjidhje për transferimin e të dhënave në kohë reale.



Nëse përdorim komandën `put`, të dhënat duhet të paketohen dhe duhet të jenë gati për ngarkim. Duke qënë se webservers gjenerojnë vazhdimisht të dhëna, kjo është një detyrë shumë e vështirë.

Për këtë duhet një zgjidhje e cila ti japë zgjidhje këtyre ‘dështimeve’ të komandës `put` dhe të transferojë të dhënat log nga gjeneruesit në HDFS në një kohë më të shkurtër dhe me më pak vonesa.

Në HDFS, skedari ekziston si një direktori dhe gjatësia e skedarit do të konsiderohet 0 (zero) derisa skedari të mbyllet. Për shembull, nëse një burim po shkruan të dhëna në HDFS dhe ndodh një shpërbërje rrejtje në mes të veprimet, pa mbyllur më parë skedarin, atëherë të dhënat e shkruajtur në skedar do të humbin.

Për këtë arsye, duhet një sistem i besueshëm, i përshtatshëm dhe mirëmbajtës për transferimin e të dhënave log në HDFS.

2.3 Zgjidhja për transferimin e të dhënave

Për të ngarkuar të dhënat stream nga burime të ndryshme në HDFS, mund të përdoren një nga mjetet e mëposhtme:

- **Facebook Scibe** – Scibe është një mjet shumë popullor, i cili përdoret për grumbullimin dhe shpërndarjen e të dhënave log. Ai është dizenuar për të shkallëzuar në një numër të madh nyjesh dhe të jetë i fortë ndaj dështime të rrjetit apo të nyjeve.
- **Apache Kafka** – Kafka është zhvilluar nga Apache SF. Ai është një ndërmjetës mesazhesh (messages broker) open-source. Duke përdorur, Kafka prurjet e të dhënave mund të trajtohen me rrjedhshmëri të lartë dhe vonesa të vogla.

- **Apache Flume** – Apache Flume është një mjet/shërbim/mekanizëm data ingestion për mbledhjen, agregimin dhe transportimin e sasive të mëdha të të dhënave streaming, si për shembull skedarët log, ngjarjet, etj nga burime të ndryshme në një data store të qëndrueshëm.

3. Arkitektura Flume

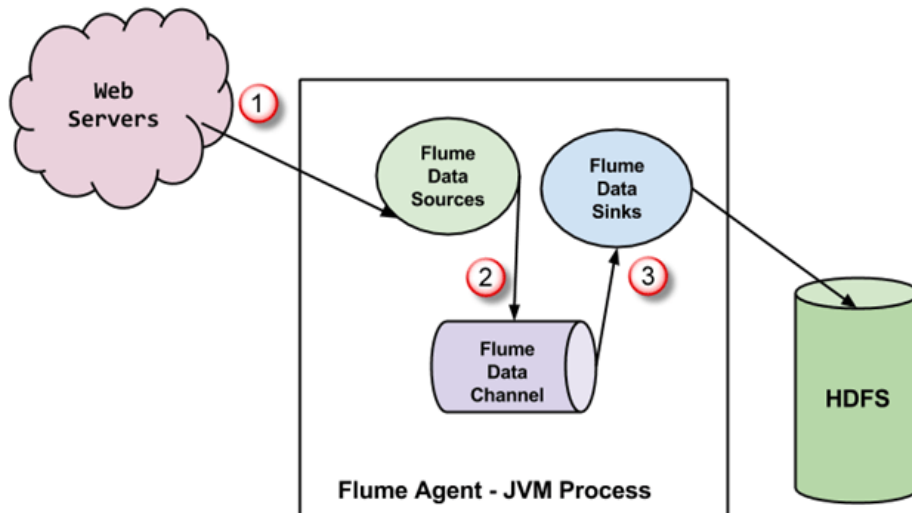





Figura 2 Arkitektura Flume

Një **agjent Flume** është një proces JVM i cili ka tre komponente:

-  *Flume Sources*
-  *Flume Channel*
-  *Flume Sink*

1. Në diagramën e ilustruar në figurën 2, ngjarjet e gjeneruara nga burimet jashtme (WebServer) konsumohen nga Burimi i të dhënave Flume (*Flume Data Source*). Burimi i jashtëm dërgon ngjarjet tek Burimi Flume në një format që njihet nga burimi i synuar.
2. Burimi Flume merr një ngjarje dhe e ruan atë në një apo më shumë kanale. Kanali shërben si një ‘depo’ ku do të ruhet ngjarja derisa të konsumohet nga Grumbulluesi (Flume Sink). Kanali mund të përdor një sistem skedari lokal për të ruajtur ngjarjet.
3. Grumbulluesi (Flume Sink) merr nga ngjarjet nga kanali, duke i fshirë ato, dhe i ruan në një ‘depo’ të jashtme, si për shembull HDFS. Mund të ketë shumë agjentë flume, në këtë rast grumbulluesi e kalon ngjarjen në agjentin flume pasardhës.

Në figurën 3, jepet një diagram për arkitekturën bazë të Flume. Ashtu siç, tregohet dhe në figurë, gjeneruesit e të dhënave (si Facebook, Twitter) gjenerojnë të dhëna të cilat mbledhen nga agjentët flume që ekzekutohen tek to. Më pas, një mbledhës të dhënash (data collector), i cili edhe ai një agjent, mbledh të dhënat nga agjentët dhe i kalon ato në një depo të qëndrueshëm, si për shembull Hbase apo HDFS.

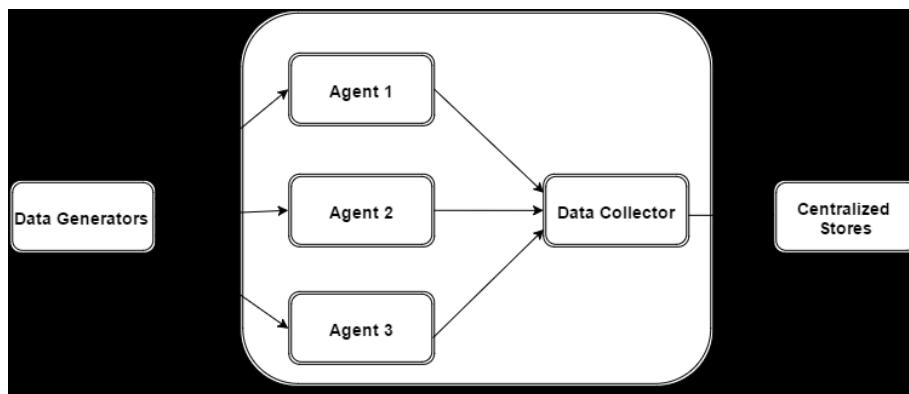


Figura 3 Arkitektura Flume

3.1 Ngjarja Flume

Një **ngjarje (event)** është njësia bazë e të dhënave që transportohen në Flume. Ngjarja përmban një sasi bytesh (njihen si Byte Payload) që duhet të transportohen nga burimi tek destinacioni dhe nga koka (header) i cili përmban informacionet e rreth sasisë së byteve. Figura 4 paraqet strukturën e një ngjarje Flume.



Figura 4 Struktura e një ngjarje Flume

3.2 Agjenti Flume

Një **agjent** është një proces i pavaruar (JVM – Java Virtual Machine) në Flume. Ai merr të dhënat (ngjarjet) nga klienti apo agjentët e tjerë dhe e përcjellë në destinacionin pasardhës (i cili mund të jetë një grumbullues ose një agjent). Flume mund të ketë më shumë se një agjent. Në figurën 5 paraqitet diagrama e një agjenti Flume, i cili përbëhet nga tre komponentë: *Burimi, Kanali dhe Grumbulluesi*. Një agjent Flume mund të ketë disa burime, kanale apo grumbulles.



Burimi (source) - Një burim është komponenti i agjentit i cili merr të dhënat nga gjeneruesi dhe i transferon ato në një ose disa kanale në formën e ngjarjeve Flume. Apache Flume suporton tipe të ndryshme burimesh, ku çdo burim merr ngjarje vetëm nga një gjenerues specifik të dhënash. Shembull: Avro source, Thrift source, twitter 1% source, etj.



Kanali (channel) - Një kanal është depo kalimtare, i cili pranon ngjarjet nga burimi dhe i ruan ato derisa të konsumohen nga grumbulluesi. Ai vepron si një urë lidhëse ndërmjet burimit dhe grumbulluesit. Këto kanale janë totalisht transaksionare dhe mund të punojnë me çdo numër burimesh dhe grumbulluesish. Shembull: JDBC channel, File system channel, Memory channel, etj.



Grumbulluesi (sink) – Një grumbullues ruan të dhënat në një depo të qëndrueshme, si Hbase dhe HDFS. Ai konsumon të dhënat (ngjarjet) nga kanali dhe i dërgon ato në destinacion. Destinacioni i një grumbulluesi mund të jetë një agjent tjetër ose depo-ja qendrore. Shembull: HDFS Sink.

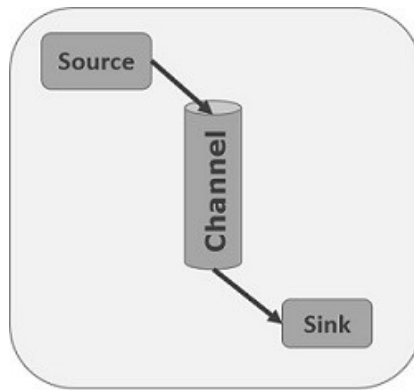


Figura 5 Struktura e Agjentit Flume

Komponentë shtesë të një Agjenti Flume

Ajo çfarë kemi përshkruar mësipër janë tre komponentët elementarë të një agjenti. Por ekzistojnë edhe disa komponentë të tjerë të cilat luajnë një rol të rëndësishëm në transferimin e ngjarjeve nga gjeneruesi i të dhënave në depot qendrore.



Interceptorë (Interceptors) – përdoren për të ndryshuar/inpektuar ngjarjet flume që po transferohen nga burimi në kanal.



Përzgjedhësit e kanalit – (Channel Selectors) – përdoren për të përcaktuar se cili kanal do të zgjidhet për të transferuar të dhënat, kjo në rastin kur agjenti përbëhet nga disa kanale. Ekzistojnë dy tipe përzgjedhësish:

- Përzgjedhësit e paracaktuar (Default channel selectors) – njihen edhe si përzgjedhësit përsëritës të kanaleve, ato kopjojnë të gjitha ngjarjet në çdo kanal.
- Përzgjedhësit e shumëfishtë (Multiplexing channel selectors) – përcaktojnë kanalin se ku do të dërgohet ngjarja, bazuar në adresën që gjendet në pjesën e header-it të ngjarjes.



Procesorët e Grumbulluesit (Sink Processors) – përdoren për të thirrur një grumbullues specifik nga grupi i grumbulluesve. Ata përdoren për të siguruar rrugë (path) të sigurta për grumbulluesit dhe ngjarjet nëpër të gjithë grumbulluesit e kanalit.

4. Apache Flume – Rrjedha e të dhënave

Flume është një framwork i cili përdoret për të ngarkuar të dhënat log në HDFS. Në përgjithësi, ngjarjet dhe të dhënat log gjenerohen nga servera në të cilët është instaluar agjent flume. Këto agjentë marrin të dhënat nga gjeneruesit e të dhënave.

Të dhënat në këto agjentë do të mbledhet nga një nyje e ndërmjetme e cila njihet si **kolektor** (Collector). Njësoj si agjentët, në Flume mund të kemi disa kolektor.

Në fund, të dhënat nga këta kolektor do të agregohet dhe do të vendoset në një depo të qëndrueshme, si Hbase apo HDFS. Në figurën 6, ilustron diagrama për rrjedhën e të dhënave në Flume.

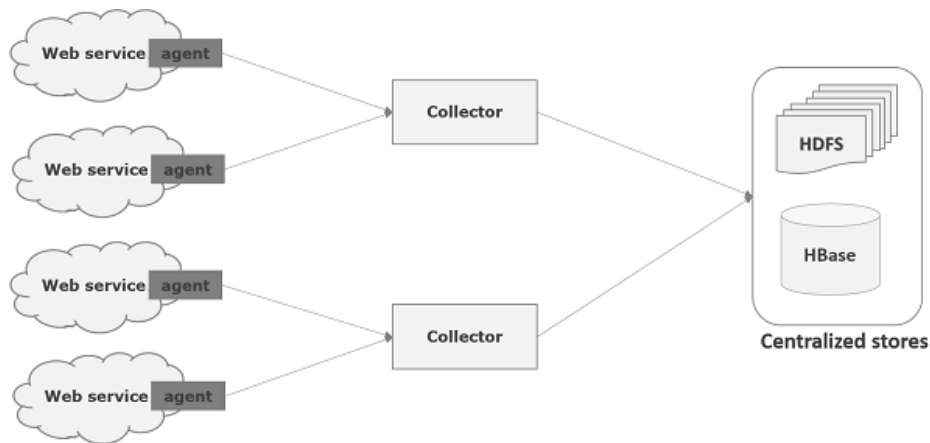




Figura 6 Rrjedhja e të dhënave në Flume

Karakteristikat e Rrjedhës së të dhënave:


 **Multi-hop Flow** – Në brendësi të Flume, mund të kemi disa agentë dhe përpara se të arrihet destinacioni final, një ngjarje kalon nëpër më shumë se një agent. Kjo njihet si *rrjedha me shumë kalime*.


 **Fan-out Flow** – rrjedha e të dhënave nga një burim në shumë kanale njihet si *rrjedha me shumë hyrje*. Kemi dy tipe:

- **Përsëritëse** – të dhënat replikohen në të gjithë kanalet e konfiguruar.
- **Shumëfishtë** – të dhënat do të dërgohen tek kanali i përzgjedhur i cili është specifikuar në header-in e ngjarjes.

Trajtimi i Dështimit

Në Flume, për çdo ngjarje realizohen dy transaksione:

 Njëra tek dërguesi










 Tjetra tek marrësi.

Dërguesi dërgon ngjarjet tek marrësi. Pasi janë marrë të dhënat, marrësi *commit* transaksionin e tij dhe i dërgon një ‘sinjal’ dërguesit. Dërguesi, pasi merr sinjalin, *commit* transaksionin e tij. *Dërguesi nuk do e bëjë commit transaksionin e tij derisa të marrë sinjal nga marrësi*, kjo për të menaxhuar më mirë rastet e dështimit të komunikimit ndërmjet dërguesi dhe marrësit.

5. Karakteristikat dhe Kufizimet e Apache Flume












Ashtu siç kemi treguar dhe më lartë, kur kemi të bëjmë me transferimin e të dhënave nga burimi tek destinacioni, Apache Flume është shërbimi më i mirë open-source për mbledhjen e të dhënave. Përveç përfitimeve, Apache Flume ka dhe disa disavantazhe/kufize të cilat do ti përmendim më poshtë.

Për mbledhjen, agregimin dhe lëvizjen e sasive të mëdha të të dhënave në mënyrë efikase në sistemin e skedarëve të shpërndarë Hadoop (HDFS) përdoret Apache Flume. Në thelb, ai është një shërbim i shpërndarë, i besueshëm dhe i disponueshëm. Për më tepër, mund të themi, se në bazë të rrjedhës së të dhënave ai ka një aritekturë të thjeshtë dhe fleksibël. Në të njëjtën mënyrë, është i fortë dhe ka një natyrë tolerante ndaj gabimeve; sidomos me mekanizma të besueshëm rekuperimin në rast dështimi. Avantazhet mund të përmbliken në:

-  Flume është i shkallëzueshëm, i besueshëm, tolerant nga gabimeve dhe i përshtatshëm për durime të ndryshme.
-  Apache Flume mund të ruaj të dhënat në magazina(stores) të qëndëruara si Hbase dhe HDFS.
-  Flume është i shkallëzueshëm horizontalisht.
-  Nëse niveli i leximit tejkalon nivelin e shkrimit, Flume siguron një rrjedhë të qëndrueshme të të dhënave ndërmjet veprimeve të leximit dhe shkrimit.
-  Flume siguron shpërndarje të besueshme të mesazheve. Transaksionet në Flume janë të bazuar në kanale, ku dy transaksione (dërguesi dhe marrësi) mirëmbahen për çdo mesazh.
-  Duke përdorur Flume, mund të ngarkohen (ingest data) nga servera të ndryshëm në Hadoop.
-  Siguron një zgjidhje të cilat janë të besueshme dhe të shpërndara, të cilat ndihmojnë në mbledhjen, grumbullimin dhe lëvizjen e sasive të mëdha të dataset-eve si Facebook, Twitter apo dhe websit-e e-commerce.
-  Ndihmon në ngarkimin në kohë reale të të dhënave log nga burime të ndryshme, si trafiku i rrjeti, mediat sociale, mesazhet/email, skedarët log, etj, në HDFS.
-  Suporton një gamë të gjërë burimesh.

5.1 Karakteristikat

Ekzistojnë shumë avantazhe kryesore të Apache Flume:

-  **Open Source** – Apache Flume është open-source, pra lehtësisht i disponueshëm.
-  **Dokumentimi** – Ekzistojnë shumë shembuj dhe modele të mira dhe sesi mund të aplikohen ato, të disponueshme në dokumentacionin e Apache Flume
-  **Vonesa** – Apache Flume ofron shkëmbimin e madh të të dhënave me një vonesë shumë të ulët; pra me shpejtësi të lartë.
-  **Konfigurimi** – Përmban instruksione shumë të thjeshta për konfigurimin dhe instalimin e tij.
-  **Rrjedha e të dhënave** – Në mjediset Hadoop, Flume punon me burimet e të dhënave streaming të cilat gjenerohen vazhdimisht, siç janë skedarët log.
-  **Routing** – Në përgjithësi, Flume shikon ngarkesën e të dhënave, si të dhënat stream apo ngjarjet; dhe gjithashtu krijon një kurs të përshtatshëm për transmetimin e të dhënave.
-  **I lirë** – Apache Flume ka një kosto të ulët për instalimin, ekzekutimin dhe mirëmbajtjen e tij.
-  **Tolerant ndaj gabimeve dhe i shkallëzueshëm** – Apache Flume është shumë i zgjeruar, i besueshëm, i disponueshëm, i shkallëzueshëm horizontalisht, si dhe i përshtatshëm për burime dhe grumbullime (sinks) të ndryshëm. Sidoqoftë, kjo ndihmonë në mbledhjen, agregimin dhe lëvizjen e të njëjës sasive të madhe të dhënash. Për shembull, Facebook, Twitter dhe website-et e-commerce.
-  **Shpërndarë** – Ai është i komponent i shpërndarë.
-  **Dorëzimi i mesazheve të besueshëm** – Ofron shpërndarje të besueshme të mesazheve. Në thelb, në Flume transaksionet janë channel-based ku dy transaksionet (dërguesi dhe marrësi) mirëmbahen për çdo mesazh.
-  **Streaming** – Na jep një zgjidhje e cila është e besueshme dhe e shpërndarë, dhe na ndihmon të ngarkojmë të dhëna nga interneti përmes burimeve të ndryshme (trafiku i rrjetit, mediat sociale, e-mail, skedarët log, etj) në HDFS.



Rrjedhje e qëndrueshme – Flume ofron një rrjedhë të qëndrueshme të të dhënave, ndërmjet veprimeve lexim/shkrim.

5.2 Kufizimet

Siç e dimë nëse ka avantazhe, patjetër që ka dhe disavantazhe. Pra, le të diskutojmë disavantazhet e Apache Flume të cilat e zvogëlojnë atë në disa aspekte. Të tilla si:



Garanci e dobët e mbajtjes së rradhës – nëse bëhet fjalë për garantimin e ruajtjes së rradhës, Apache Flume është shumë i dobët.



Dublikimi – Në shumë skenarë, Flume nuk garanton që mesazhi që merret është unik. Sidoqoftë, ekziston mundësia që mesazhet të dublikohen.



Shkallëzueshmëri e ulët – Ekziston një mundësi e vogël që për një ndërmarrje, madhësia e paisjeve tipike të një Flume të jetë pak e ndërlikuar, dhe në shumicën e rasteve, të rezultojë në gabim. Prandaj, thuhet se aspekti i shkallëzimit të Flume është shpesh i ulët.



Çështjet e besueshmërisë – Në rastet kur zgjedhja e depos ruajtëse nuk është zgjedhur me ‘mençuri’ duke marrë të gjithë faktorët ndikues, atëherë vihen në diskutim shkallëzueshmëria dhe besueshmëria e Apache Flume.

6. Aplikimet e Apache Flume

Siç e dimë kur bëhet fjalë për trajtimin e të dhënave që rrjedhin në/nga bazat e të dhënave relacionale dhe lëvizjen e shpejtë e të dhënave të pastrukturuara mund të përdorim Apache Flume. Gjithësesi, ka dhe shumë raste të tjera kur mund të përdorim Apache Flume. Më poshtë është dhënë një listë e të gjithë rasteve të mundshme të përdorimit të apache Flume:



Kur duam të marrim të dhëna nga burime të ndryshme dhe ti ruajmë ato në sistemin Hadoop, në mund të përdorim Apache Flume.



Sa herë që kemi nevojë për të trajtuar të dhëna me shpejtësi të lartë dhe vëllim të lartë në sistemin Hadoop, atëherë përdorim Apache Flume

- Ndhmon gjithashtu në shpërndarjen e besueshme të të dhënave në destinacion.



Kur shpejtësia dhe volumi i të dhënave rritet, Flume duke përdorur si zgjidhje shkallëzimin, ai mund të ekzekutohet shumë thjesht edhe nëq shtojmë disa paisje të tjera.

- Pa ndalur për asnjë moment, Flume konfiguron në mënyrë dinamike komponentët e ndryshëm të arkitekturës.
-



Kur duam të përdorim të dhënat në kohë reale, atëherë përdorim Apache Flume.








Përdorim Flume, kur duam të mbledhim të dhëna log nga servera të ndryshëm dhe ti ruajmë ato në një depo të qëndrueshme, HDFS, Hbase.



Me ndihmën e Flume, mund të mbledhim të dhëna nga servera të ndryshëm si në kohë reale ashtu dhe formën e batch-eve.



Mund të importojmë të volume shumë të mëdha të dhënave të gjeneruara dhe të analizuara në kohë reale nga faqet e internetit të mediave sociale, si Facebook dhe Twitter dhe faqeve të internetit të tregtisë elektronike (e-commerce) si Amazone dhe Flipkart.

-  Duke përdorur Flume, është e mundur që të mbliidhen të dhëna nga një bashkësi e madhe burimesh dhe më pas të kalohen në destinacione të ndryshme.
-  Flume suporton Multi-hop flows, fan-in fan-out flows dhe routing kontekstual.
-  Përdorim Apache Flume, kur kemi disa aplikacione web, të cilat gjenerojnë log-e ose kur duam të ngarkojmë log-et në HDFS me një shpejtësi të lartë.
-  Për maskimin e të dhënave apo filtrimin e tyre Flume mund të jetë shumë i dobishëm.
-  Flumë mund të shkallëzohet horizontalisht.

7. Apache Flume

Në Apache Flume, Burimi Flume (Flume Source) është mjeti i cili merr/pranon të dhënat nga gjeneruesit e të dhënave, dhe gjithashtu i transferon ato në një ose disa kanale si ngjarje Flume (Flume events). Në figurën 7, tregohen tipet e ndryshme të burimeve flume të disponueshëm:

- Avro Flume,
- Thrift,
- Exec,
- JMS,
- Spooling Directory,
- Flume Kafka,
- NetCat TCP,
- NetCat UDP,
- Sequence Generator,
- syslog në Flume

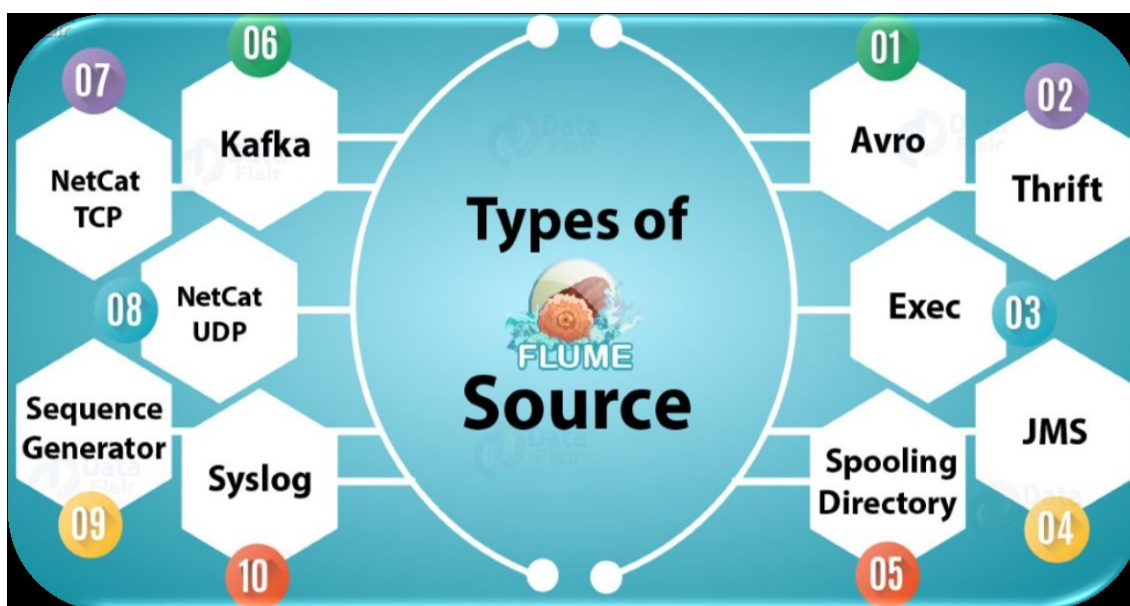
















Figura 7 Burimet Flume

Komponentët Flume Sink përdoren për të ruajtur të dhënat në data store të qëndësuar. Data store në Hadoop janë Hbase dhe HDFS. Ky komponent konsumon ngjarjet nga kanalet dhe i ‘dorëzon’ ato në destinacion i cili mund të jetë një agjent tjetër apo një data store. Në figurën 8, tregohen tipet e ndryshme të grumbulluesve që përdoren nga një agjent Flume:

-  HDFS Sink,
-  Hive Sink,
-  Logger Sink,
-  Thrift Sink,
-  IRC Sink,
-  File Roll Sink,
-  HBase Sink,
-  MorphlineSolrSink,
-  ElasticSearchSink,
-  Kite Dataset Sink,
-  Flume Kafka Sink,
-  HTTP Sink,
-  Custom Sink in flume,
-  Apache Flume Avro Sink

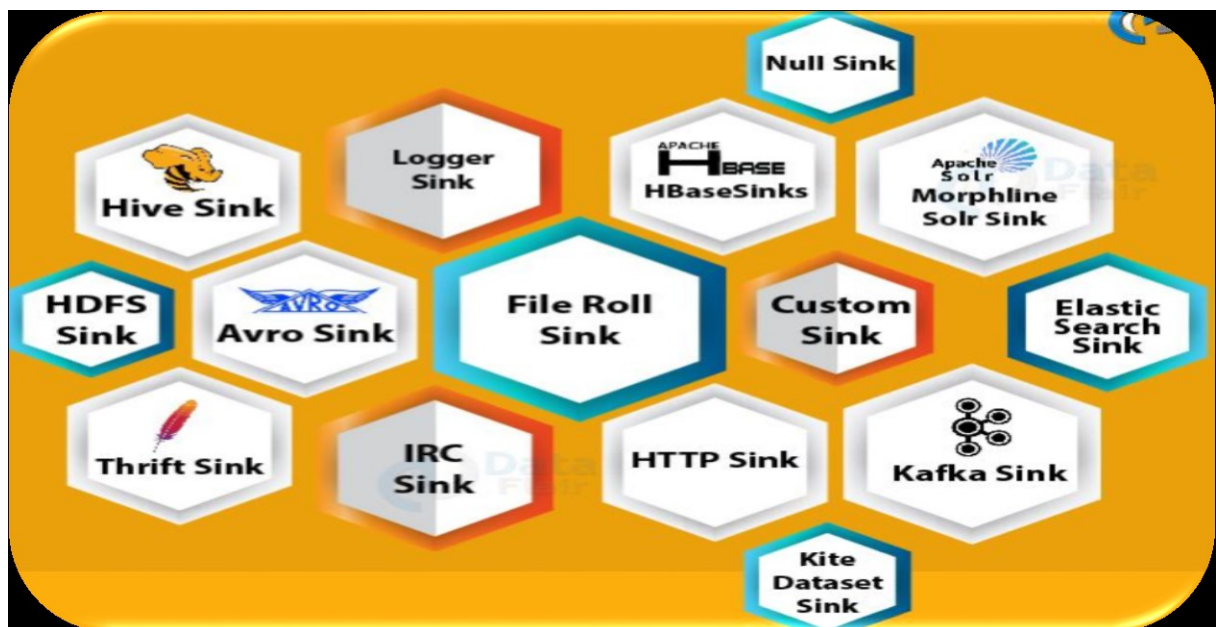









Figura 8 Kanalet Flume

Kanalet Apache Flume janë depo kalimtare, të cilat marrin ngjarjet nga burimet i ruan ato derisa konsumohen nga grumbulluesi. Me fjalë të tjera mund të themi se kanali luan rolin e një ‘ure’ ndërmjet burimeve dhe grumbulluesve. Këto kanale mund të punojnë më tipe të ndryshme burimesh dhe grumbulluesve Flume. Kanalet Flume kanë natyrë transaksionale. Në figurën 9, tregohen tipet e ndryshme të kanaleve që përdoren nga një agjent Flume:

-  Memory Channel
-  File Channel

-  JDBC Channel
-  Custom Channel
-  Kafka Channel
-  Spillable Memory Channel
-  Pseudo Transaction Channel

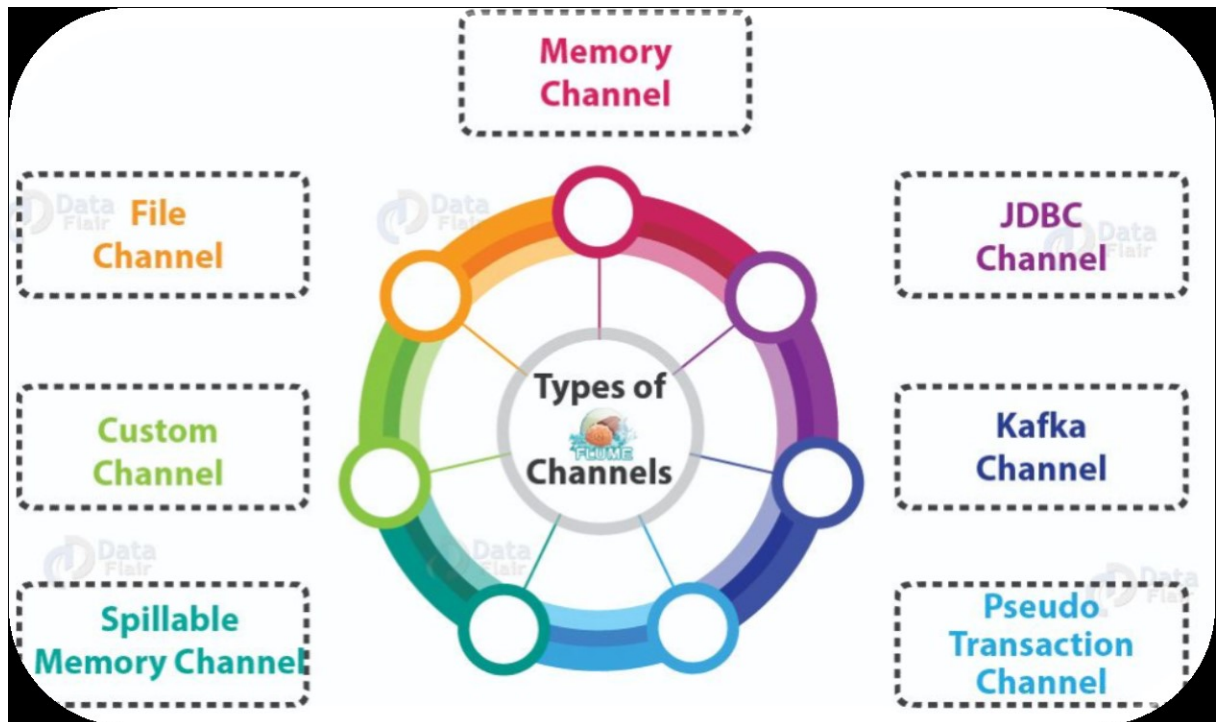





Figura 9 Kanalet Flume

8. Setup -Shembull

Konfigurimet e Agjentit Flume ruhen në një **skedar tekst lokal konfigurimi**. Ky skedar është formatin JAVA. Konfigurimet për një ose disa agjentë ruhen në të njëjtin skedar. Skedari përfshin karakteristikat (properties) për çdo Burim, Grumbullim, Kanal dhe mënyrën sesi ato lidhen me njëra-tjetrën për transportimin e të dhënave.

Për çdo komponentë (burim, kanal, apo grumbullim) specifikohen:

-  Emrin
-  Tipin
-  Karakteristikat specifike të tipit dhe Inicializimin e tyre

Për shembull:

- Burimi Avro:
 - Emri: r1 (një emër identifikues i burimit, i vendosur nga përdoruesi)
 - Tipi: Avro

- Specifikimet për Hostname/Adrese IP dhe Numrin e portës të burimit nga do merren të dhënat.
- Kanali i memories, përveç emrit dhe tipit duhet të ketë të specifikuar kapacitetin maksimal të dhënave.
- HDFS Sink përveç emrit dhe tipit, kërkon specifikimet për file system URI dhe direktorinë ku do ruhen skedarët

Agjenti duhet të dijë se cilat komponentë duhet të ngarkojnë dhe si janë të lidhur në mënyrë që të krijojë rrjedhën e të dhënave. Kjo bëhet duke renditur emrat e secilit prej burimeve, sinks dhe kanalet në agjent, dhe më pas duke specifikuar kanalin lidhës për secilin sink dhe burim.

Për shembull:

- Një agjent kalon ngjarje nga një burim Avro –avroWeb -në HDFS sink - hdfs-cluster1- përmes një kanali skedarësh - file-channel -.
- Skedari i konfigurimit do të përmbajë emrat e këtyre komponenteve dhe kanalin e skedarëve si një kanal të përbashkët për të dy burimin avroWeb dhe hdfs-cluster1sink.

Një agjent startohet duke përdorur komandën **flume-ng**, duke specifikuar emrin e agjentit, direktorinë e skedarit të konfigurimit dhe emrin e këtij skedari:

```
flume-ng agent -n $agent_name -c conf -f conf/flume- conf.properties.template
```

Mëposhtë, paraqitet një skedar konfigurimi test, example.conf, për inicizimin e një agjenti flume në single-node, ku është specifikuar që do të ketë burim *Netcat*, grumbullues *Logger* dhe kanal *Memory*

```
# example.conf: A single-node Flume configuration

# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1

# Describe/configure the source
a1.sources.r1.type = netcat
a1.sources.r1.bind = localhost
a1.sources.r1.port = 44444

# Describe the sink
a1.sinks.k1.type = logger

# Use a channel which buffers events in memory
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

Për të startuar agjentin flume me këto konfigurime, ekzekutohet komanda e mëposhtme, në Windows Shell:

```
flume-ng agent --conf conf --conf-file example.conf --name a1 -Dflume.root.logger=INFO,console
```


Në një terminal/paisje tjetër, bëjmë **telnet port 44444** dhe dërgojmë një ngjarje Flume, si mëposhtë.

```
$ telnet localhost 44444
Trying 127.0.0.1...
Connected to localhost.localdomain (127.0.0.1).
Escape character is '^]'.
Hello world! <ENTER>
OK
```

Pëgjigja që marrim nga agjenti Flume është:

```
INFO source.NetcatSource: Source starting
INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChannelImpl[/127.0.0.1:44444]
INFO sink.LoggerSink: Event: { headers:{} body: 48 65 6C 6C 6F 20 77 6F 72 6C 64 21 0D      Hello world!. }
```