AWS Cloud ပေါ်မှာ Data Lake တွေဆောက်ကြမယ်

S3, Data Lakes, Data Analytics

Data က ကျွန်တော်တို့ နေ့စဉ် ဘဝမှာ ထိတွေ့နေရတဲ့ အရာပါပဲ။ website တစ်ခုခုက article ရှာဖတ် လိုက်တာလည်း data ယူသုံးလိုက်တာပေါ့။ computer ထဲသိမ်းထားတဲ့ excel fileတွေ ထုတ်ကြည့်တာလည်း data အသုံးချလိုက်တာပါပဲ။

Data တွေကို အကြမ်းဖျင်းခွဲရင်

1. Structured data (excel file data တွေ၊ SQL database တွေ)
2. Semi Structured data (XML files တွေ၊ နေ့တိုင်းသုံးတဲ့ email တွေပါတယ်)
3. Unstructured data (Video၊ image၊ audio file တွေပေါ့)

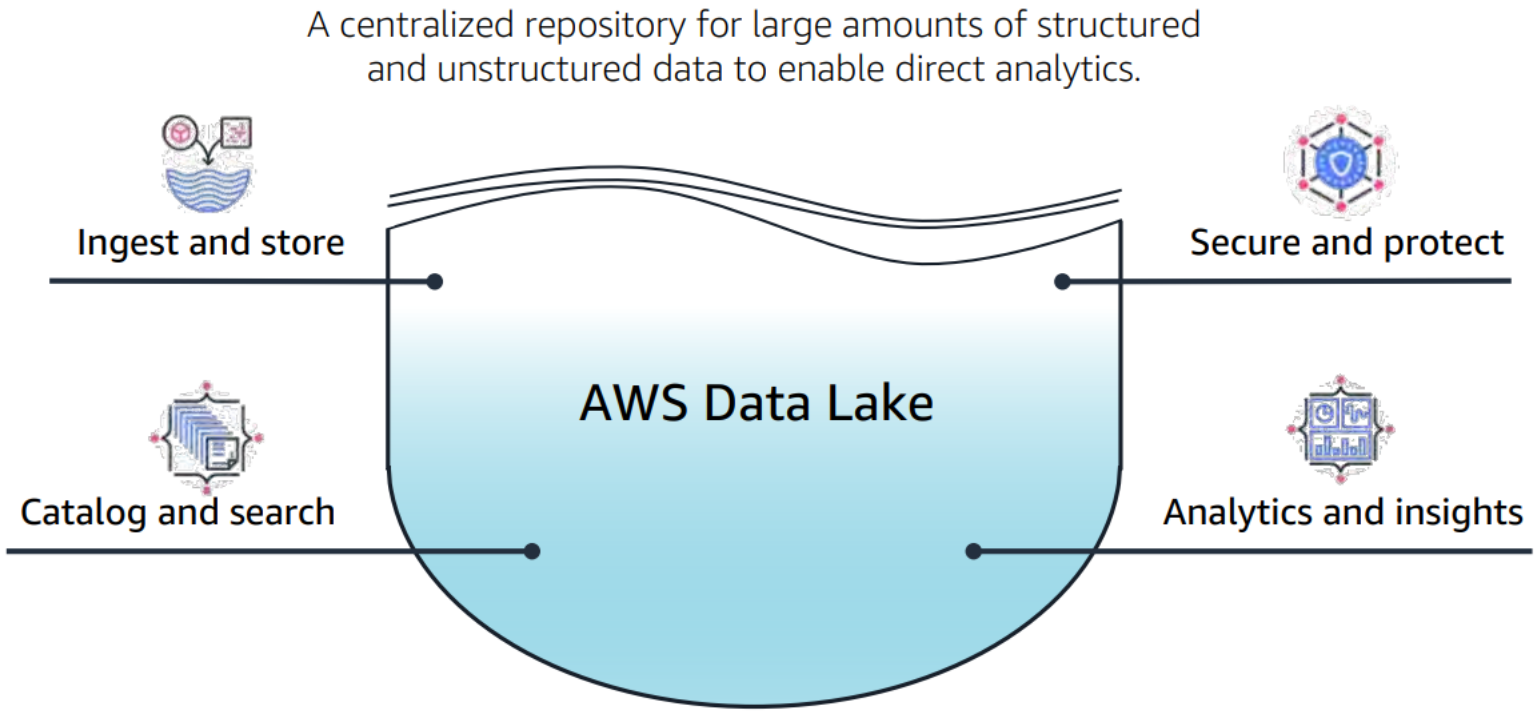
AWS Data lakes အကြောင်းဆိုတော့ data အကြောင်းအကျယ်မချဲ့တော့ပါဘူး။ အိုကေ။ data တွေကတော့ ဟုတ်ပြီ။ ပိုကြီးလာတဲ့ data တွေ၊ ဥပမာ movie streaming service ဆို terabytes, petabytes နဲ့ချီတဲ့ data တွေဖြစ်လာပြီ။ အဲ့ကျ ဒီdata တွေကို server ကြီးတွေနဲ့ သိမ်းရပါတယ်။

Infra အကြောင်း ဆိုတာနဲ့ on premises ရယ် cloud ရယ် ဆိုပြီး စကွဲပါပြီ။

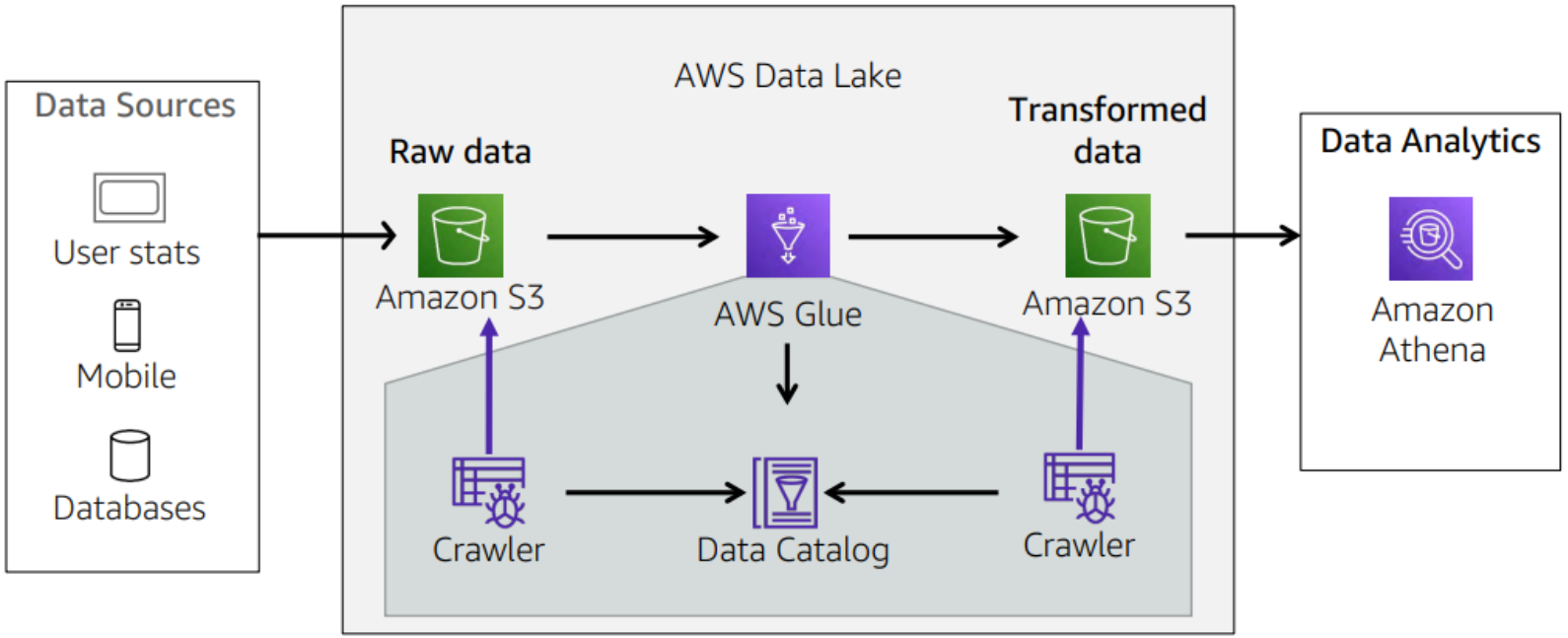
On premises မှာက Data silos ဆိုတာတွေထားကြတယ်။ Department တစ်ခု မှာ ဒီ department manage လုပ်တဲ့ server ပေါ့။ data silos တွေက department တိုင်းမှာ ရှိနေနိုင်သလို တစ်ခုတည်းမှာ data silos တွေများကြီးရှိနေနိုင်တယ်။ ပြောချင်ကတော့ data တွေက centralized မဖြစ်တော့ အကုန်စုပြီး process ရတဲ့ task တွေ efficient မဖြစ်ဘူး၊ time taken ကြာတာပေါ့။

Cloud ပေါ်မှာတော့ data lake ဆိုပြီး centralized repository၊ ပင်မ သိမ်းတဲ့နေရာဆိုပြီး ဆောက်ထားလို့ရတယ်။ Data lake ဆောက်မယ်ဆို AWS ပေါ်မှာဆို S3 ရှိတယ်။ စျေးအသက်သာဆုံး storage server ပဲ။ Data lake အဖြစ် S3 ကိုသုံးရတာက object storage ဖြစ်တာကြောင့် flexible ဖြစ်တာရယ်၊ structured, unstructured ကြိုက်တာ သိမ်းလို့ရတယ်။ bucket versioning, lifecycle policy တွေရှိတော့ data ပေါ်မူတည်ပြီး align လုပ်ရတာ အဆင်ပြေတယ်။

ခုလို ပြောလို့ on premise data silo တွေ မသုံးရဘူးလား။ လုံးဝမဟုတ်ဘူး။ business credential data တွေဆို on premise, own server နဲ့ သိမ်းရမယ်လေ။ အပြင်ကို leak လို့မရဘူး။ cloud ရွေးမလား၊ on premise လားဆိုတာက ဒါမျိုး factor တွေကျ စဉ်းစားရမယ်။ ခု article ကတော့ data analytics pipeline ဆောက်ရင် firstly data တွေ စုပြီး collect ဖို့ data lake ဆောက်တဲ့အကြောင်းပြောထားတာမလို့ အဲ့factor ဖယ်ထားပါတယ်ဗျ။ ;D



S3 ကို သုံးပြီး Data lake ဆောက်ပြီးရင် ဘာတွေလုပ်ရလဲ။ Data တွေ collect ဖို့ store ဖို့ လိုရမယ်။ အဲ့တာကို security ကောင်းအောင် encryption တွေဘာတွေထားတာတွေလုပ်လို့ရတယ်။ Data analytics အကြောင်းကတော့ နောက်article မှာဆက်မှာဖြစ်လို့ data processing, data querying solution တွေ ဒီမှာ မစတော့ပါဘူး။



ဒီ diagram ကတော့ S3 ကို data lake အဖြစ်သုံးထားတဲ့ data analytic architecture ပုံစံလေးပါ။ I will share this lab with explanation soon. ဒီ article မှာ ပြောထားတဲ့အတိုင်း data sources တွေကနေ အားလုံးကို စုပြီး သိမ်းတဲ့ centralized repository အဖြစ် S3 ကို သုံးတယ်။ အဲ့ကနေမှာ data integration tool ဖြစ်တဲ့ Glue သုံးပြီး S3 ထဲပြန်ထည့်ပါတယ်။ S3 ကနေပဲ Querying tool Athena သုံးပြီး data တွေ ပြန်ကြည့်တာပေါ့။ နောက် article မှာ Glue services, Athena တွေသုံးတာပြောပါမယ်။ data warehouse အဖြစ် Redshift ကိုသုံးနိုင်တာရယ်၊ data warehouse vs data lake ရယ်၊ data analytics အကြောင်း နည်းနည်း စီအကျယ်ချဲ့ပါ့မယ်ခဗျ။

Thank you