

EXAMENSARBETE Evaluating ClickHouse as a Big Data Processing Solution for IoT-Telemetry**STUDENTER** Adrian Göransson, Oskar Wändesjö**HANDLEDARE** Markus Borg (LTH), Anton Friberg (Axis)**EXAMINATOR** Alma Orucevic-Alagic (LTH)

ClickHouse för skalbar analys och lagring av massiva mängder data

POPULÄRVETENSKAPLIG SAMMANFATTNING **Adrian Göransson, Oskar Wändesjö**

Insamling och analys av stora mängder data för Business Intelligence är idag inget ovanligt. Insamlingen av data har ökat i aggressiv takt, och i samband med detta har utmaningar att skala upp sin analys- och lagringskapacitet vuxit. Detta arbete har undersökt hur ClickHouse, en kolumnbaserad databas, kan ta sig an dessa utmaningar.

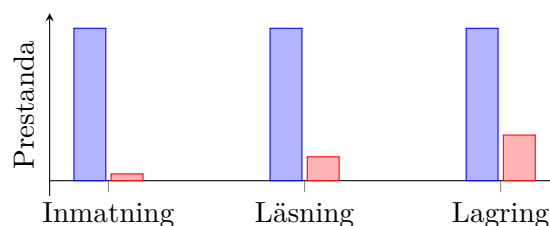
För att få insyn och kunna basera beslut på stora mängder data krävs det att man har en välplanerad infrastruktur som låter en lagra och hämta data på ett effektivt sätt. Relationsdatabasen är en väl beprövad lagringsmodell som funnits sedan 70-talet, där data sparas som rader i olika tabeller. Modellen erbjuder strikta garantier för integriteten av inmatad data och är väl lämpad för många och små transaktioner där en blandning av operationer som läser, skapar eller uppdaterar data görs. I större system, där datan som matas in inte kommer att förändras, utan bara läsas, exempelvis metrikdata för analys, kan dock dessa strikta garantier istället leda till prestandaproblem.

De senaste 15 åren har flertalet databaser som gått från den radbaserade tabellmodellen utvecklats i olika former, till exempel dokumentdatabaser och grafdatabaser. De erbjuder oftast inte samma strikta garantier som relationsdatabasen, och kan därför prestera bra även vid uppskalning.

ClickHouse är en kolumnbaserad databas som ursprungligen utvecklats av Yandex, Rysslands största internetbolag, men som hösten 2021 avknoppades till ett amerikanskt bolag i ClickHouse Inc. ClickHouse är specialanpassat för höga skriv- och läshastigheter och utlovar mycket effektiv lagring optimerad för många typer av data.

I detta arbete testade vi ClickHouse hos Axis, ett företag som samlar in stora mängder metrikdata från en bred flotta med IoT-enheter (Internet of Things). Analysavdelningen har stött på utmaningarna att skala upp sin analys, som primärt beror på deras val av datalagringslösning.

Vi utformade en experimentell testsvit för att undersöka främst hur ClickHouse stod sig mot dagens lösning i form av sökdatabasen Elasticsearch. Vi testade inmatning och läsning av metrikdata, samt hur effektivt de olika subjekten lagrade data.



Figur 1: Relativa prestandaresultat i olika tester för ClickHouse (blå), och Elasticsearch (röd).

Våra resultat visade att ClickHouse i snitt var 22 och 6,4 gånger snabbare än Elasticsearch i inmatning respektive läsning samt 3,3 gånger mer effektiv på att lagra datan.