

Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačních technologií

BAKALÁŘSKÁ PRÁCE
DATOVÁ INTEGRACE NEZPRACOVANÝCH DAT Z IOT

Autor: Ondřej Bednář

Vedoucí práce: Ing. Jan Masner, Ph.D

2023

Motivace

- Narůstající trend využití IoT
- Velký výběr DI řešení
- Využití dat v oblasti Machine Learning

Cíle práce

Hlavní cíl:

Výběr ETL nástroje pro integraci dat ze senzorů používaných v zemědělství

Dílčí cíle:

- charakteristika dostupných řešení
- vytvoření AHP modelu
- výběr řešení pro následnou implementaci
- implementace těchto řešení a porovnání jejich výkonu

Metodika

Přehled odborných informačních zdrojů

- charakteristika používaných ETL nástrojů
- výběr vhodných kritérií

Sestavení rozhodovacího modelu AHP

- stanovení vah

Vyhodnocení modelu

- výběr dvou nejlepších řešení

Implementace řešení

- provedení zátěžového testu

Teoretická část

Charakteristika dat

- co odlišuje IoT od jiných datových zdrojů

Součásti řešení datové integrace

- rozdílný přístup Batch ETL x Stream ETL

Základní kategorizace nástrojů

Přehled obdobných řešení

Vlastní práce

Představení případu užití

Data sázecích strojů

- Soustava více samostatných secích zařízení
- Senzory osiva
- Hloubkové senzory
- Pneumatické senzory
- Geopoziční data
- Připojení k internetu

Integrace dat v reálném čase

Cloudové prostředí

Middleware IoT gateway

Logický datový model



Zdroj: Deere & Company

Vlastní práce

Sestavení modelu AHP

Kritéria

- Dvě stěžejní
- Jedno doplňkové

Váhy

- Zohledněný daný use case

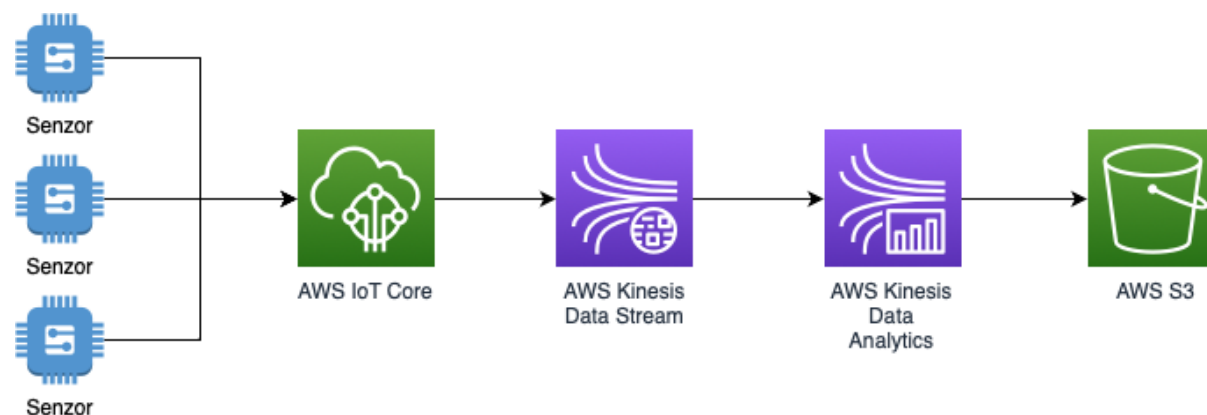
Alternativy

- Azure Stream Analytics
- AWS Kinesis
- Google Dataflow

		Název kritéria	Váha ve skupině	Váha normovaná
Provozní	0,4	Efektivita kódu	0,122	0,049
		Podporované jazyky	0,227	0,091
		Dostupné API	0,227	0,091
		Aplikace třetích stran	0,423	0,169
Technická	0,4	Nativní připojení	0,249	0,100
		Real-time připojení	0,558	0,223
		Latence	0,096	0,039
		Propustnost	0,096	0,039
Ostatní	0,2	Rozsah dokumentace	0,143	0,029
		Požizovací náklady	0,429	0,086
		Provozní náklady	0,429	0,086

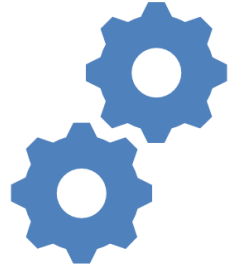
Vlastní práce

- Implementace nejvhodnějších řešení na základě modelu AHP
- Integrace v cloudovém prostředí
- Úplný popis implementace
- Ukázky konkrétních výstupů z použitých nástrojů



Zdroj: vlastní práce

Vlastní práce



Návrh vhodného testu

Sledovaná latence a využití výpočetních zdrojů

Často používané ETL transformace

Simulace odběru dat uživatelem



Testování implementovaných řešení

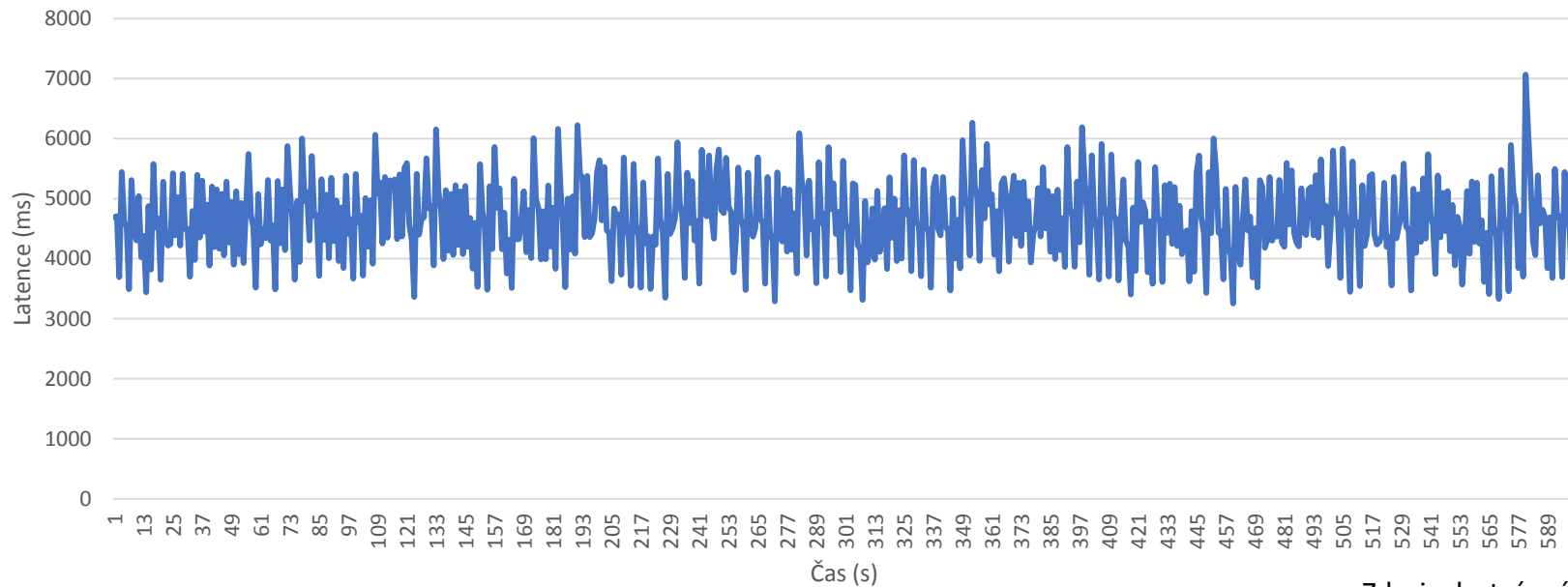
Výsledky

Přehledné srovnání nástrojů

Kritérium	Váha	Sub-kritérium	Alternativa					
			ASA	AWSK	GCD	ASA	AWSK	GCD
			Váha ve skupině			Váha normovaná		
Provozní	0,4	Efektivita kódu	0,637	0,258	0,105	0,255	0,103	0,042
		Podporované jazyky	0,500	0,250	0,250	0,200	0,100	0,100
		Dostupné API	0,600	0,200	0,200	0,240	0,080	0,080
		Poskytované transformace	0,637	0,258	0,105	0,255	0,103	0,042
Technická	0,4	Nativní připojení	0,200	0,600	0,200	0,080	0,240	0,080
		Real-time připojení	0,088	0,669	0,243	0,035	0,268	0,097
		Latence	0,135	0,822	0,333	0,054	0,329	0,133
		Propustnost	0,333	0,333	0,333	0,133	0,133	0,133
Ostatní	0,2	Rozsah dokumentace	0,200	0,400	0,400	0,040	0,080	0,080
		Pořizovací náklady	0,333	0,333	0,333	0,067	0,067	0,067
		Provozní náklady	0,455	0,091	0,455	0,091	0,018	0,091
					Suma	1,450	1,521	0,945

Výsledky

- Vyhodnocení zátěžových testů
- Vizualizace výsledků
- Výstupy z procesu implementace



Zdroj: vlastní práce

Závěr

Provedena charakteristika dostupných řešení

Sestaven model AHP, nejvhodnější alternativy AWS Kinesis, Azure Stream Analytics

Nástroje zhodnoceny na základě zkušeností z implementace

Nástroje podrobeny zátěžovému testu, výsledky zanesené do grafů

Oba nástroje vyhodnoceny jako vhodné pro daný případ užití

Doporučen ETL nástroje – AWS Kinesis (nižší latence, více možností využití)

Děkuji za pozornost