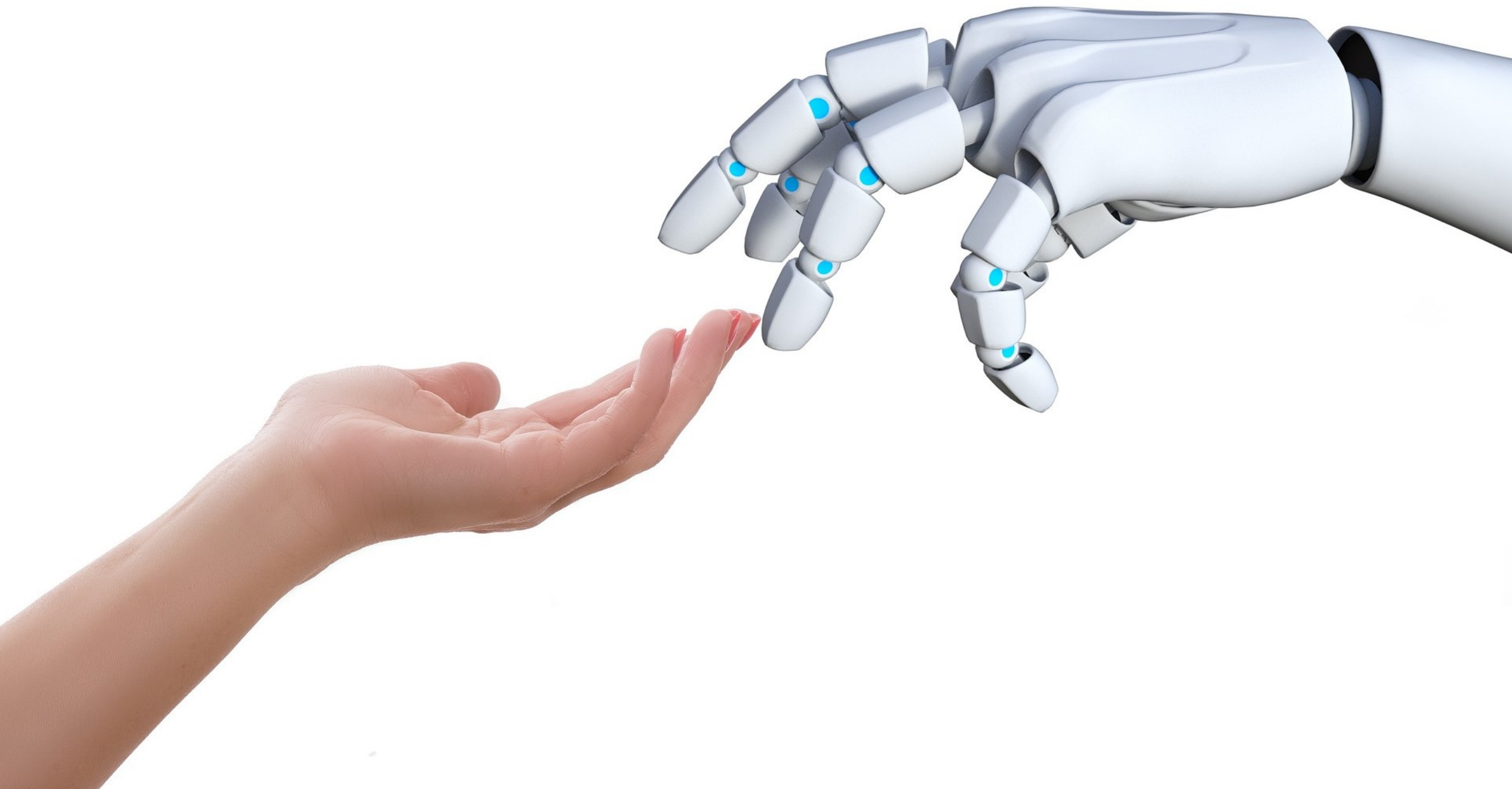


Technologické trendy v datové analytice na Univerzitě Pardubice

Digitální transformace univerzit

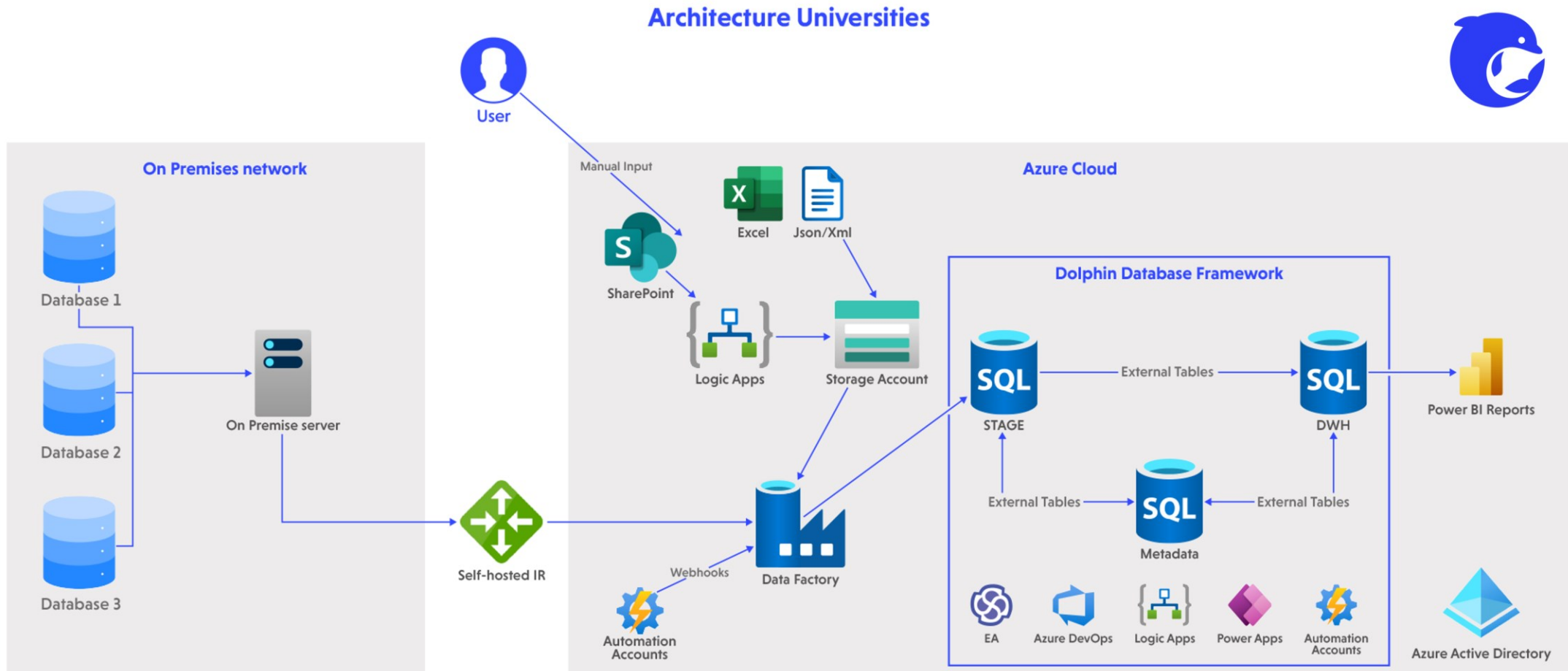
10. – 11. 9. 2024



Aktuální řešení

- MS Azure
 - Dvě relační databáze – STAGE a DWH
 - Data Factory (ELT služba, orchestrační nástroj)
 - Objekty nejsou vytvářeny manuálně
 - Framework zvaný „flašinet“
 - Zakládá DB objekty, vytváří transformační pipeliney, generuje základní skelet transformačních procedur, ...
 - Zajišťuje také dodržování osvědčených metodik a *best practices*.
 - Dvě prostředí
 - Vývojové vs. produkční
-

Architektura DWH 1.0 z vrtulníku



If it works, don't touch it.



- Infrastruktura

- UPCE jako první univerzita – implementace sady doporučení a osvědčených *best-practice* známých jako tzv. **landing zóna**
 - Přehledné cloudové prostředí,
 - Správa prostředí pomocí kódu (Infrastructure as Code, IaC),
 - Systematické řízení životního cyklu jednotlivých služeb
 - Transparentní přidělování oprávnění zaměstnancům i dodavatelům

- DWH historicky první velká věc v tenantu

- Výchozí subskripce, vlastní networking, ...
 - → snaha standardizovat s ohledem na pravidla landing zóny
 - Zabezpečení
-

Motivace – II

- Vývojové a produkční prostředí
 - Z historických důvodů
 - Vývojové prostředí v tenantu implementačního partnera
 - Produkční prostředí v tenantu univerzity
 - Ve vývojovém prostředí
 - Vývoj nad omezenou množinou dat
 - Někdy výhoda, často nevýhoda
 - Snaha dostat obě prostředí do tenantu UPCE
 - Snazší samovývoj
 - Jasnější exit strategie (cloud, dodavatel, ...)
 - Co největší odladění již během vývoje
-

Motivace – III

- Flašinet
 - Řízený metadaty (modelem)
 - Struktury zdrojového systému; cílové struktury DWH
 - Rychlé zpracování standardních scénářů
 - Rizika
 - Změna zdrojových systémů; chybně zadaná data; ...
 - Potenciální důsledky
 - Dočasný výpadek celého ETL procesu
 - Ztráta části historie (historizace na straně STAGE)
 - Cíl: vyšší odolnost proti změnám; zrychlení 1. fáze vývoje (prototypy)
-

Motivace – IV

- Modernizace technologického stacku s ohledem na stávající možnosti
 - Využít nové možnosti Data Factory
 - Zbavujeme se limitů DB
 - Limit 4 TB, aktuálně využito 3,5 TB
 - Standardizované úložné formáty
 - Bez vazby na konkrétní dodavatele
 - Příprava na nové architektonické přístupy k datové analytice
-

Motivace – V

- Relační databáze mají řadu předností
 - A některé nevýhody
 - Velké objemy nestrukturovaných dat
 - Komplexní datové typy
- Snaha:
 - Připravit si půdu pro AI analýzy
 - Nestrukturovaných dat
 - Rozsáhlých textů
 - ...

Storing JSON in a SQL Database



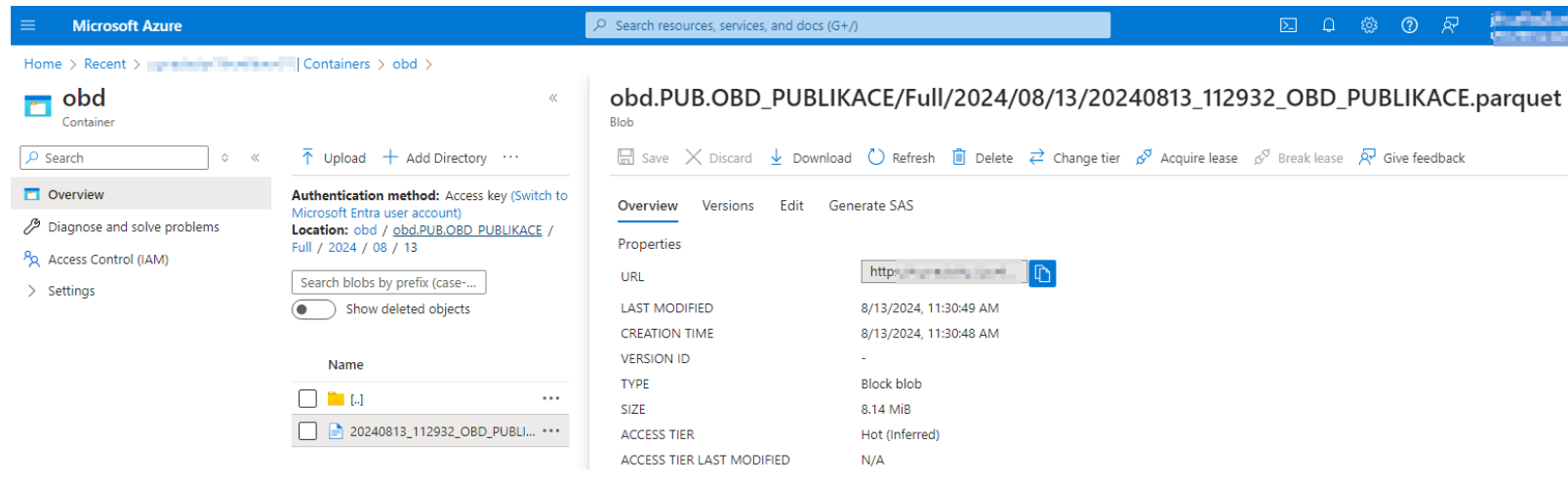


JAK??



STAGE vrstva

- Zrušení STAGE ve formě relační databáze
- Implementace architektury Data Lake
 - Ukládání nejen strukturovaných dat, ale „čehokoli“
 - Důležitá správná metodika organizace dat

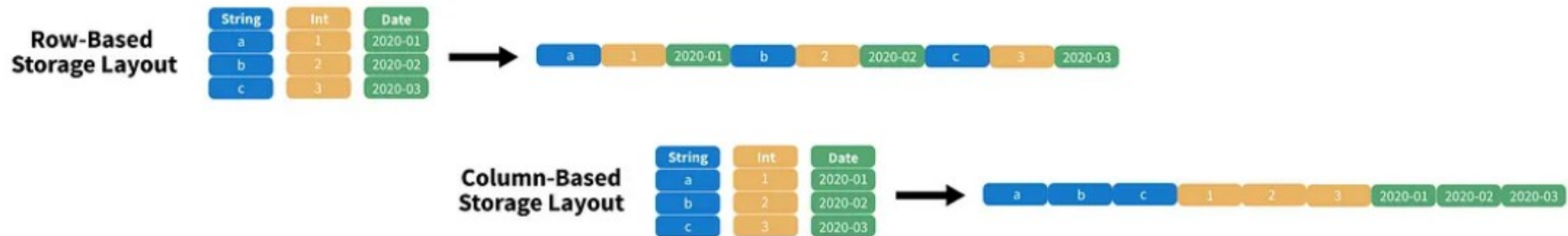


The screenshot displays the Microsoft Azure portal interface. On the left, the 'obd' container is selected, showing its overview and a list of blobs. The main pane displays the details for the blob 'obd.PUB.OBD_PUBLIKACE/Full/2024/08/13/20240813_112932_OBD_PUBLIKACE.parquet'. The blob's properties are listed below.

Properties	
URL	http://obd.blob.core.windows.net/obd.PUB.OBD_PUBLIKACE/Full/2024/08/13/20240813_112932_OBD_PUBLIKACE.parquet
LAST MODIFIED	8/13/2024, 11:30:49 AM
CREATION TIME	8/13/2024, 11:30:48 AM
VERSION ID	-
TYPE	Block blob
SIZE	8.14 MiB
ACCESS TIER	Hot (Inferred)
ACCESS TIER LAST MODIFIED	N/A

STAGE vrstva

- Pro strukturovaná data (nejen relační) → **Apache Parquet**
 - Open source formát, de facto průmyslový standard
 - Souborový formát, sloupcově orientovaný
 - Sloupcová komprese



<https://www.linkedin.com/pulse/what-parquet-file-format-why-special-sai-prasad-padhy-u1rzc/>

STAGE vrstva

[illegible]

Preview

New Folder

Select All

Copy

Paste

Clone

More

ult)

<< 2024 > 08 > 13

Filter by prefix (case-sensitive)

Access Tier

LIKACE.parquet

Hot (inferred)

Access Tier Last Modified

ZDROJ_NAZEV	ZDROJ_ZK...	KONANI_...	VYDANI	VYDAVAT...	
Chefferie traditionnelle et l'Etat en Af...		Niamey, Niger		Niamey	
Antropologični izsledovanja				Sofia	4
The Intellectual Legacy of Ernest Gell...				Moscow	
The Intellectual Legacy of Ernest Gell...				Moscow	
Sci. Pap. Univ. Pardubice Ser. C				Pardubice	
The Intellectual Legacy of Ernest Gell...				Moscow	
Sociologický časopis/Czech Sociologi...	null			Praha	3
		Dolní Rožň			
null		Luháčovice			
Český časopis historický	null			Praha	1
Český časopis historický	null			Praha	1
Sci. Pap. Univ. Pardubice Ser. C				Pardubice	2
Sci. Pap. Univ. Pardubice Ser. C				Pardubice	2
Sci. Pap. Univ. Pardubice Ser. C				Pardubice	2
Sci. Pap. Univ. Pardubice Ser. C				Pardubice	2
VIII. International Meeting on Choline...		Perugia, Italy			
Proceedings of 6th International Con...		Kouty nad Des...		Pardubice	6
Proceedings of 31st International Co...		Tatranské Matl...	31	Bratislava	3
Reaction Kinetics and Catalysis Letters	React. Kinet. C...			Budapešť	8
11th Nordic Symposium on Catalysis		Oulu, Finland			
7th Pannonian International Symposi...	null	Srní, Česká Re...	null	null	r
3rd EFCATS School on Catalysis	null	Ustron, Poland	null	null	r
Sborník Univerzity Pardubice	null	Pardubice	1	Pardubice	r
Podnikání v praxi			1	Praha	3
Podnikání v praxi			1	Praha	4
Podnikání v praxi			1	Praha	4
Podnikání v praxi			1	Praha	4
Podnikání v praxi			1	Praha	4

tem selected)

<<

<

1

>

>>

Load more

Showing 1 to 100 of 42873 items

<<

<

1

>

>>

Load more

STAGE vrstva

- Dynamická struktura parquet
 - Nemusí být dopředu modelován
 - Při změně primárních datových struktur se přizpůsobí („samoseť“)
 - „Spadnou“ až navazující operace
 - Nenastane výpadek uložených dat (historie zůstává zaznamenána)
 - Sjednocení diverzity datových typů (např. Oracle vs. MS SQL)
-

STAGE vrstva

- Na parquet lze napojit i PowerBI
 - Aktivace načítání nových dat do parquetu rychlá
 - → předpoklad rychlé tvorby prototypů
 - Nutná správná governance

Datafactory

- Využití nových možností
 - Dynamický metadaty řízený nahrávací proces
 - Zadáno „co“ má být nahráno
 - O zbytek (např. lokalizaci) se postará sám
 - Možnost asynchronního loadu
 - Možnost inkrementálního loadu
-

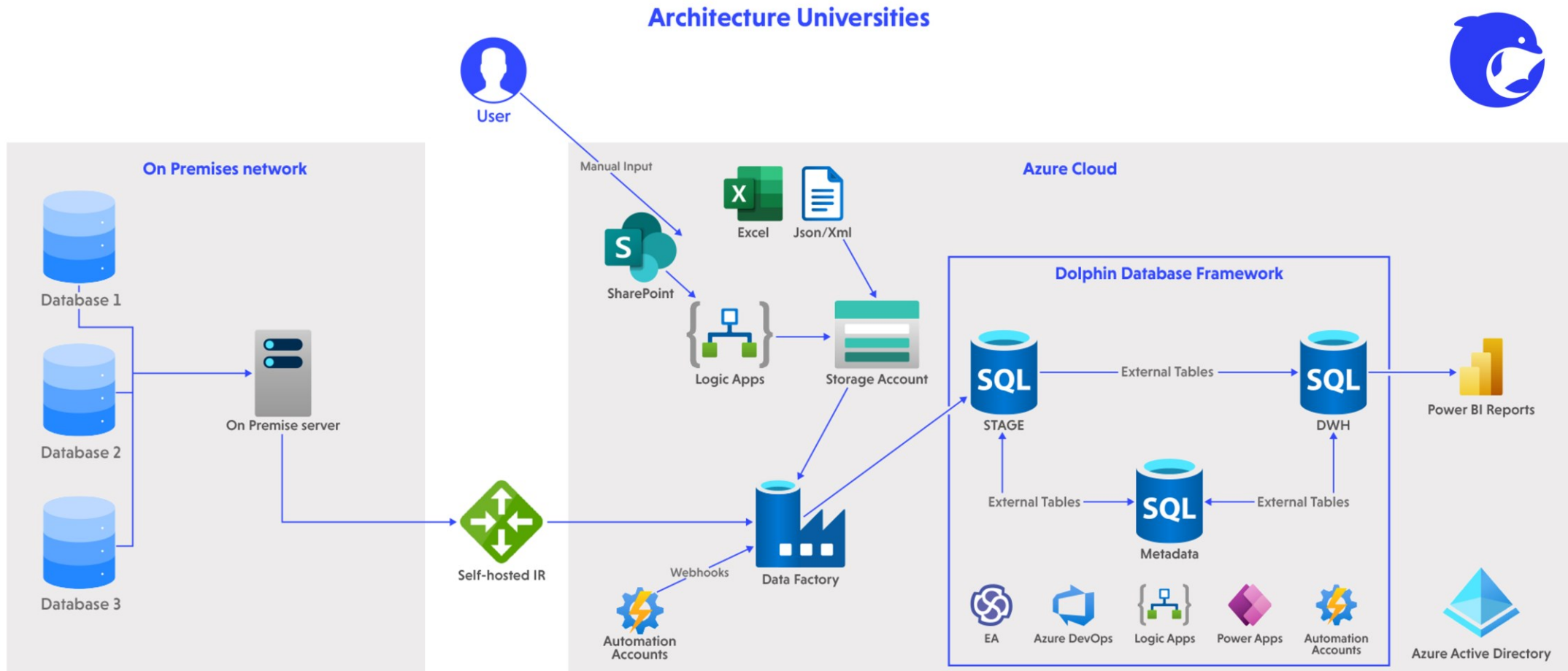
DWH vrstva

- Zůstane ve formě relační databáze
 - Zachování know-how a logiky transformací z DWH 1.0
 - Nebudou transformovat tabulky ze STAGE databáze
 - Ale „tabulky“ z parquet souborů
 - Ideální cílový stav
 - Staré sestavy fungující nad novým DWH
 - Vracející ideálně stejné výsledky 😊
-

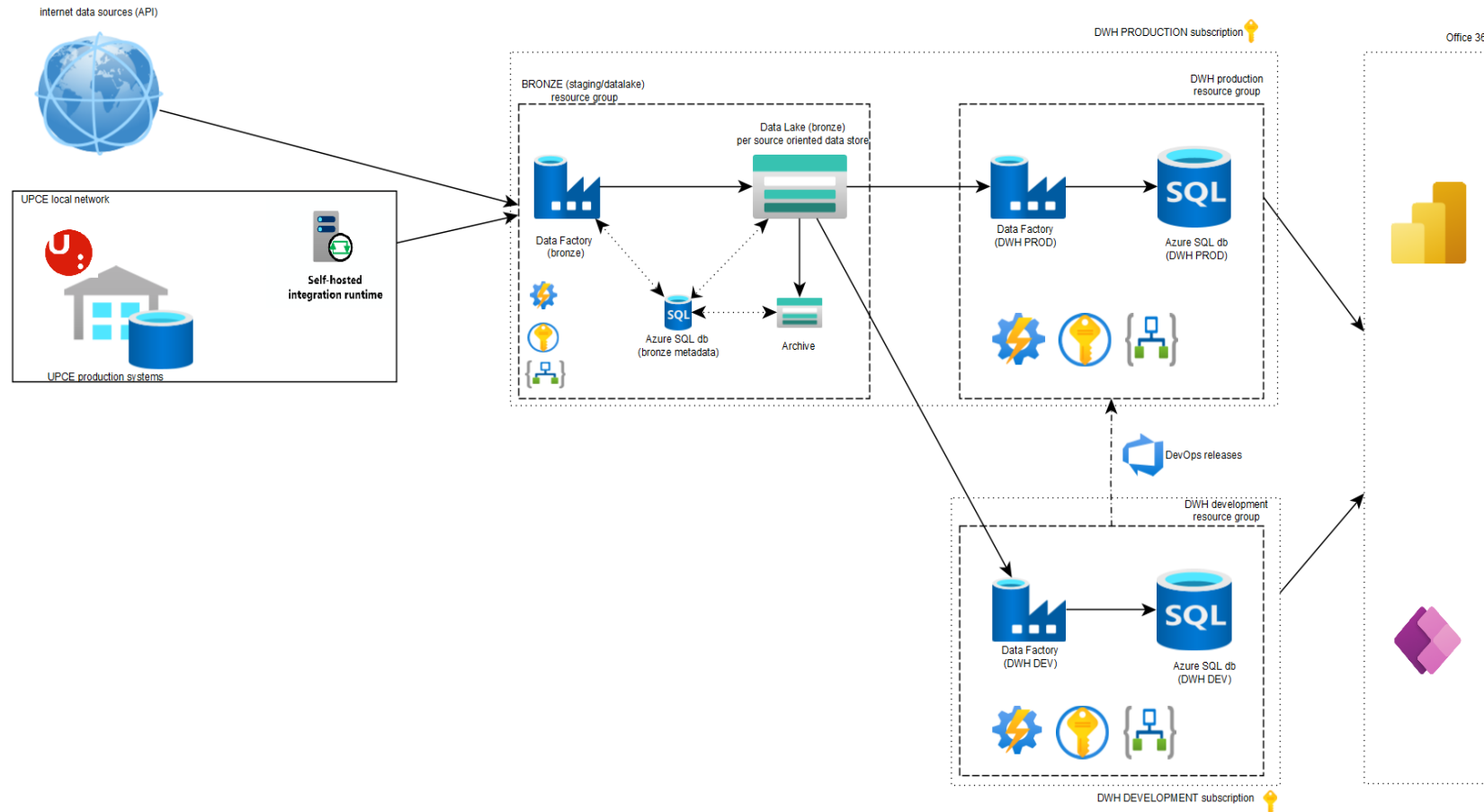
Prostředí

- Sdílená STAGE pro vývojové i produkční prostředí
 - Vývoj nad maximální množinou dat
 - Minimalizace překvapení po nasazení na produkci
 - Vývojové prostředí
 - De facto sloučené vývojové a testovací prostředí
 - = kompromis mezi chtěným a možným
 - Příprava na případný automatický deployment
-

Architektura DWH 1.0 z vrtulníku



Architektura DWH 2.0 z vrtulníku



DĚKUJEME ZA POZORNOST

Jiří Šafra <jiri.safra@upce.cz>

Ondřej Gřešák <ondrej.gresak@dolphinconsulting.cz>



**UNIVERZITA
PARDUBICE**
